A Survey of Generative Adversarial Networks for Synthesizing Structured Electronic Health Records

GHADEER O. GHOSHEH, University of Oxford, United Kingdom
JIN LI, Nanjing University of Information Science and Technology (NUIST), China
TINGTING ZHU, University of Oxford, United Kingdom

Electronic Health Records (EHRs) are a valuable asset to facilitate clinical research and point of care applications; however, many challenges such as data privacy concerns impede its optimal utilization. Deep generative models, particularly Generative Adversarial Networks (GANs), show great promise in generating synthetic EHR data by learning underlying data distributions while achieving excellent performance and addressing these challenges. This work aims to survey the major developments in various applications of GANs for EHRs and provides an overview of the proposed methodologies. For this purpose, we combine perspectives from healthcare applications and machine learning techniques in terms of source datasets and the fidelity and privacy evaluation of the generated synthetic datasets. We also compile a list of the metrics and datasets used by the reviewed works, which can be utilized as benchmarks for future research in the field. We conclude by discussing challenges in GANs for EHRs development and proposing recommended practices. We hope that this work motivates novel research development directions in the intersection of healthcare and machine learning.

CCS Concepts: • **Computing methodologies** → **Machine learning**; • **Applied computing** → **Health informatics;**

Additional Key Words and Phrases: Generative models, generative adversarial networks, electronic health records, synthetic data

ACM Reference format:
Ghadeer O. Ghosheh, Jin Li, and Tingting Zhu. 2024. A Survey of Generative Adversarial Networks for Synthesizing Structured Electronic Health Records. *ACM Comput. Surv.* 56, 6, Article 147 (January 2024), 34 pages.
https://doi.org/10.1145/3636424

1 INTRODUCTION

Since the early 2010s, machine learning models have proven to have a high potential for supporting medical applications by using data collected in **electronic health records (EHRs)** [149, 177]. Hospitals and medical providers are increasingly adopting and deploying EHR systems. In the U.S. alone, 84% of hospitals adopted EHR systems as of 2015, which is a ninefold increase since

JL was funded by the National Natural Science Foundation of China. TZ was supported by the RAEng Engineering for Development Research Fellowship and the National Institute for Health Research Oxford Biomedical Research Centre.

Authors' addresses: G. O. Ghosheh and T. Zhu, University of Oxford, Oxford, United Kingdom; e-mails: {ghadeer.ghosheh, tingting.zhu}@eng.ox.ac.uk; J. Li, Nanjing University of Information Science and Technology (NUIST), Nanjing, China.

0360-0300/2024/01-ART147 $15.00
https://doi.org/10.1145/3636424

2008 [79]. The widespread recording of structured EHRs is paving the way for research opportunities in healthcare applications, such as patient-stratification [152], drug repurposing [33], public health surveillance [16], as well as the novel discovery of disease mechanisms and correlations as seen in the early COVID-19 applications [44]. EHRs also provide a valuable asset to develop data-driven and patient-specific clinical decision support systems for diagnostic, prognostic, healthcare cost containment and workflow improvement applications [110, 154, 160]. However, the full utilization of the wealth of the EHR data in such applications is impeded by several challenges, including data sharing and privacy concerns [101], where data protection guidelines and regulations such as the Health Insurance Portability and Accountability Act [61] in the United States and the **General Data Protection Regulation (GDPR)** [170] in Europe have detailed controlling measures that prevent direct access to much of the data for patient privacy purposes. Other data-specific challenges that make EHR processing burdensome include class imbalance [151], data missingness [118], noise [103], heterogeneity [42], and irregular sampling [159]. To mitigate these challenges, deep generative models have been proposed to generate synthetic data [31], notably **variational autoencoders (VAE)** [105], and **Generative Adversarial Networks (GANs)** [69].

In this article, we review GANs for EHR applications, which is a fast-emerging yet understudied application of deep generative models. There exist several reviews and surveys related to GANs evaluation [147], GANs applications for medical imaging [96], and observational health data [64]. However, in this survey, we focus on GANs for structured EHRs and their applications, evaluation, and challenges, which serve as a basis for a reading audience with diverse backgrounds. Other related reviews include the work of Brophy et al. [19] and Zhang et al. [199] where their focuses were time-series data in general and not necessarily EHRs. One of the most related works is that of Hernandez et al. [80], where the authors reviewed the works of GANs for EHRs. Although the authors described the applications of GANs for EHRs, they only focused on a narrow aspect of GANs—the "generation" functionality. In comparison, our work reviews a broad range of applications of GAN for EHRs and is not restricted to the native generative functionality. Examples of applications included in our work are treatment effects estimation, semi-supervised learning, missing value imputation and privacy preservation. Furthermore, Hernandez et al. grouped the evaluation metrics into three categories: resemblance evaluation, utility evaluation, and privacy evaluation, respectively. While useful, the grouping is limited to metrics only used to evaluate the native data-generation task and does not include those for other applications. In addition, the resemblance evaluation category grouped metrics on a high level without describing the differences in the purpose, or mathematical implications of each metric. Other related works such as that of El Emam et al. [56] examined the impact of utility metrics on the predictive performance of synthetic data generated using GANs but did not consider the privacy and qualitative assessment that are presented in our work.

To this end, our work extends the previous reviews of GANs for EHRs to include applications such as imputation and treatment effects. Overall, we provide a comprehensive and up-to-date review of the current works and group them based on their target application in healthcare, not only for generating synthetic samples but also for mitigating many of the data challenges in the EHR domain. To the best of our knowledge, this is the first work to present a comprehensive grouping and provide a more detailed comparison of the various metrics used with mathematical formulation, examples from the literature and a discussion on the utility of such metrics for tabular and time-series EHRs. We also discuss several open-ended challenges and themes to motivate new research directions in both the computational and healthcare fields. Relevant literature was identified by searching Google Scholar using the following keywords and keyword combinations, (1) "*GAN*" AND "*EHR*," (2) "*synthetic health data*," and (3) "*GAN*" AND "*Health*" up until January

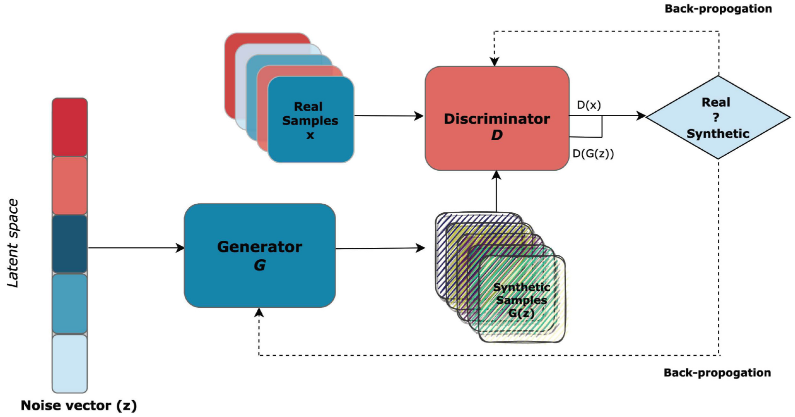


Fig. 1. An overview of the architecture of GANs showing the function of both the *generator* and *discriminator* neural networks. The generator takes an noise vector **z** as input and outputs the synthetic data. The discriminator is trained to distinguish between the real and synthetic data. Both G and D are then fine-tuned by back-propagation.

2022. We then filtered out papers that used generative models other than GANs, also those used biomedical signals collected from wearable data as well as duplicates

The outline of the article is as follows. In Section 2, we briefly review the working principles and architecture of GANs and provide an overview of EHR data types in Section 3. We then survey the research papers that used GANs for various EHR applications, in Section 4. We discuss and curate a list of commonly used evaluation metrics in Section 5, along with the most commonly used data sources in the literature in Section 6. We conclude by discussing challenges as well as future directions of GANs for EHRs in Section 7.

2 GENERATIVE ADVERSARIAL NETWORKS

2.1 Principles and Architecture

Since the introduction of GANs in 2014 [69], they have shown great potential in generating realistic data for various applications. The working principle of GANs essentially involves the training of a pair of deep neural networks in competition with each other [43]. The first neural network, the *generator* G, takes a noise vector $\mathbf{z}$ from latent space as an input and generates the synthetic samples $G(\mathbf{z})$ [43], while the other neural network, the *discriminator*, D is given both the real $\mathbf{x}$ and generated samples $G(\mathbf{z})$ and is trained to discriminate between the real and synthetic ones [69]. The discriminator outputs a vector of probability predictions of whether the inputted samples were real or synthetic. Both the generator and discriminator are fine-tuned using the discriminator's output via back-propagation as shown in Figure 1. The training involves both finding the parameters of a discriminator that maximize its classification accuracy and finding the parameters of a generator that minimize the discriminator's ability to tell the real and synthetic samples apart [69]. In other words, the objective loss function for the original GANs is as follows:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{x}}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}}\left[\log(1 - D(G(\mathbf{z})))\right].$$

2.2 GAN Variants

The initial results of GANs were promising [69], which motivated researchers to propose modifications and adaptations for specific tasks and applications. Notably, Reference [130] proposed the Conditional Generative Adversarial Net, which generated data by conditioning the GAN on a selected variable or label **y**, where **y** is fed to the generator and discriminator as a part of the input. Another important work is deep convolutional GAN, which utilized a pair of deep convolutional networks for each of the G and D [141]. Around the same time, an Information Maximizing Generative Adversarial Network was proposed to provide additional interpretability where semantic meaning was introduced to the variables in the latent space [32]. **Recurrent GAN (RCGAN)** [57] extended the original GAN model to generate sequential data by using **recurrent neural networks (RNNs)** for EHR applications, motivating several GAN applications for time-series data. Other important works include *CycleGAN* [205] and *STARGAN* [39], which were adapted to allow for domain translation, and diversity sensitive conditional GAN, which regularizes the generator to produce diverse outputs [184].

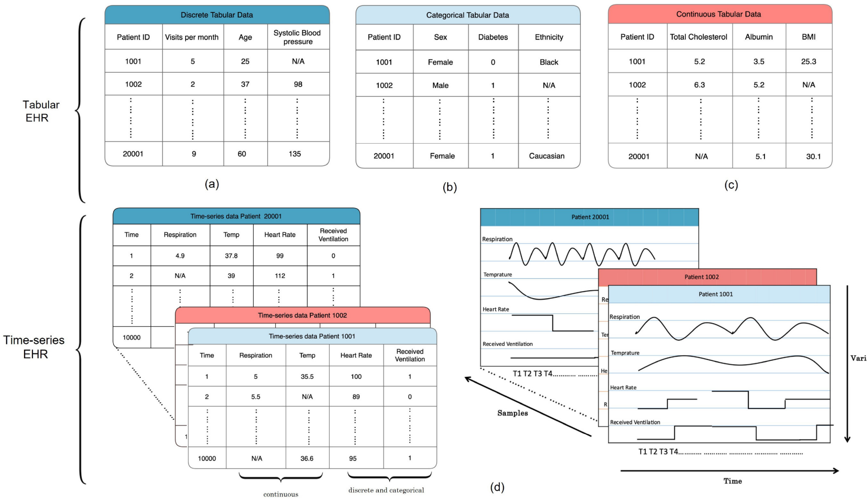
2.3 GAN Training Challenges

Despite their high potential, training GANs involves many challenges, notably mode collapse. Mode collapse refers to the case where the generator maps different inputs to a small set of synthetic outputs rather than producing diverse outputs that reflect the input [69]. Another challenge is vanishing gradients [4], where the discriminator is performing very well and not providing useful information to improve the generator training, leading the generator's gradient to vanish [4]. To address these challenges, some architectures and modifications to the loss function were proposed as seen in **Wasserstein Generative Adversarial Networks (WGAN)**. WGAN modified the loss function to improve GAN training stability by using the Wasserstein distance metric to measure the distribution similarity of real and synthetic data [5, 72]. Other modifications were minibatch discrimination [68, 148], minibatch averaging [36], unrolled GANs [128], and noise injection [148]. Notwithstanding the advantages of GANs, improving GAN training stability remains one of the bottlenecks in scaling GAN applications in real-world settings.

2.4 Related Deep Generative Models

While GANs are considered as one of the prominent deep generative models, they belong to a family of models that demonstrate strong performance in different applications. For example, likelihood-based models such as VAEs [105] and diffusion models [85] are commonly used in various deep generative applications [140, 143, 186] where diffusion models recently outperformed GANs [52]. Both VAEs and diffusion models rely on mapping the data to a latent space representation where the generation process involves learning a transformation from the latent space to the observational data. Despite their similarities such as having lower-bound-based loss functions, VAEs tend to learn compressed embeddings in latent space while diffusion models' embeddings are noisy augmentation of the original data. Other related models include transformers that were introduced in 2017 [169]. Transformers rely on self-attention mechanisms and positional encoding for learning sequence-to-sequence mapping. While often used for machine translation, text summarization, and other natural language processing tasks, transformer-based models were proposed for other generalized tasks in imaging [102] and tabular data generation [9].

All of the aforementioned models perform well in various generative tasks, but they present limitations. VAEs are very good at learning a compressed version of the data; however, they often generate noisy and blurry outputs with compromised quality [129]. Diffusion models, however, are inefficient in training and tend to generate less private synthetic data compared to other deep generative models [26]. Compared to adversarial models such as GANs, likelihood-based models such as VAEs and diffusion models, transformers require large datasets to train and often have a large number of parameters to optimise to achieve stable performance, and hence more inherently complex.



Discrete Tabular Data			
Patient ID	Visits per month	Age	Systolic Blood pressure
1001	5	25	N/A
1002	2	37	98
⋮	⋮	⋮	⋮
20001	9	60	135

Categorical Tabular Data			
Patient ID	Sex	Diabetes	Ethnicity
1001	Female	0	Black
1002	Male	1	N/A
⋮	⋮	⋮	⋮
20001	Female	1	Caucasian

Continuous Tabular Data			
Patient ID	Total Cholesterol	Albumin	BMI
1001	5.2	3.5	25.3
1002	6.3	5.2	N/A
⋮	⋮	⋮	⋮
20001	N/A	5.1	30.1

Time-series data Patient 20001				
Time	Respiration	Temp	Heart Rate	Received Ventilation
1	4.9	37.8	99	0
2	N/A	39	112	1
⋮	⋮	⋮	⋮	⋮
10000				

Time-series data Patient 1002

Time-series data Patient 1001				
Time	Respiration	Temp	Heart Rate	Received Ventilation
1	5	35.5	100	1
2	5.5	N/A	89	0
⋮	⋮	⋮	⋮	⋮
10000	N/A	36.6	95	1

Fig. 2. The two main types of EHR data, tabular and time series, are shown in their various forms. Discrete, categorical, and continuous tabular data are shown in (a), (b), and (c), respectively. Time-series data are shown in (d), where the record is shown on the left and a corresponding plot of the data is shown on the right.

3 STRUCTURED EHRS: DATA TYPES AND CLINICAL SETTINGS

In medical practice, medical staff use EHRs to record and capture various forms of data about a patient during an encounter. Like paper records, EHRs store data such as hospitalization information and patient-level information such as demographics, comorbidities, medical history, vital signs, laboratory tests, prescribed medication, administered interventions, diagnosis, and clinical outcomes [16]. The nature of each of these kinds of data differs, which results in multiple types of EHR data. While there are similarities between physiological signal data such as **photoplethysmography (PPG)** and **electrocardiogram (ECG)** and structured EHRs, the purpose and nature of the recording is often different. For the sake of this work, we focus on structured EHRs collected in clinical settings that are often presented in either tabular or time-series formats, as shown in Figure 2. Tabular data store information that presents a representation of the patient's encounter such as demographic features, aggregated mean, or a one time measurement of vital signs, where each sample has one value for each feature. Time-series data, however, present a record of data points indexed in time order, which might be used to present disease progression over time as seen in longitudinal data [81] or even short-term records as seen in vital signs [157]. The variables recorded in each of the two data types can be discrete, categorical, or continuous. Discrete variables represent values that can be obtained by counting and stored as integers such as age or number of visits per month, as seen in Figure 2(a) and (d). Categorical variables, however, are used when there is a finite number of categories such as sex or ethnicity, as shown in as seen in Figure 2(b) and (d). Last, continuous variables are variables whose value is obtained by measurement and is not limited to whole numbers. Examples of continuous variables can be seen in many laboratory tests and vital signs such as albumin, body temperature, and total cholesterol, as shown in Figure 2(c) and (d). It is worth noting that different EHR data types usually coexist in the same patient record. For example, a patient might have both tabular and time-series data recorded for the same visit. This heterogeneous nature of EHRs often results in complexity in terms of its analysis, modeling, and use for machine learning purposes [42, 177].

EHR data can be recorded in different settings and stages of a patient encounter or observation. During a hospital visit, a patient encounter can be classified as either inpatient or outpatient, where the first requires hospitalization and admission, while the latter does not. For an inpatient encounter, a patient could go through various units within the same facility, which depends on the clinical status [195], availability of human and material resources [51], or hospital capacity [59]. At the beginning of a hospital presentation, patients can be presented to the emergency wards where initial diagnosis and interventions take place [162], where the focus is to admit and then triage the patient based on the medical need. In the general inpatient-wards, patients get regular laboratory tests, vital sign checks, treatment administration, and other required procedures as requested by the doctor. Patients who deteriorate or those whose cases require higher care are admitted to the **Intensive Care Unit (ICU)**, where the data tend to be frequently collected as the patient is under close monitoring. Data collected in ICUs are usually referred to as critical care data [91]. The other types of EHR data are those of outpatient encounters, where the data collected are for patients who were not admitted to the hospital, as seen in the case of specialist consultations [78] and visits to general practitioners [81]. The nature of outpatient data varies across countries, depending on the availability of primary care and the need for referrals to get a specialist consultation.

4 APPLICATION OF GANS FOR STRUCTURED EHRS

The applications of GANs in the medical domain are very diverse, specifically in medical imaging. For instance, GANs have been used for various radiology tasks that ranged from data augmentation to data segmentation and denoising [120, 189, 201]. However, there is much less work on using GANs to generate realistic structured healthcare data such as EHRs. The lag in the use of GANs for EHR data can be attributed to the many data challenges, such as complexity, heterogeneity, and missingness [177]. In comparison with other data modalities such as images and text, which can be intuitively and visually evaluated for realism, assessing the quality of the generated EHR data is difficult. In Table 1, we summarise major works that used GANs for EHR applications and group them based on their target application. The main groups are (1) generation of diverse types of EHRs, (2) semi-supervised learning and data augmentation, (3) imputation of missingness, (4) treatment effect estimation, and (5) privacy preservation. The works are reviewed in terms of the used models, task, dataset size, open-access code and data, as well as evaluation components used to assess the quality of the synthetic data.

4.1 Generation of Diverse Types of EHRs

In the following subsections, we describe GAN-based works that generated different types of EHR data, tabular and time series, in Sections 4.1.1 and 4.1.2, respectively. We also survey papers that attempted to explore heterogeneity aspects in either tabular or time-series EHRs in Section 4.1.3.

4.1.1 Generating Tabular EHRs. The early GANs for EHRs works focused on generating structured discrete tabular EHRs such as diagnosis and billing ICD codes. For example, *medGAN* was one of the first GANs architectures to address the incompatibility of the original GANs to generate tabular EHRs with binary or discrete count features [37]. The authors' model incorporated an autoencoder to learn the salient features of discrete variables in tabular EHRs, which assists GANs in learning the distribution of multi-label discrete binary and count features. Building on the success of *medGAN* for generating discrete data, *medWGAN* and *medBGAN* were proposed based on **Wasserstein GAN with gradient penalty (WGAN-GP)** [72] and **boundary-seeking GANs (BGAN)** [84], respectively. The authors' major contribution was in the area of improving the quality of generated data of that generated by the original *medGAN* [10]. In *MC-medGAN*, the authors proposed adaptations to medGAN to allow for better representation of multi-categorical

Table 1. Summary of the Various Uses of GANs for EHRs and Comparison of Target Application, Evaluation Measures, Medical Datasets, and Open Access

		Problem		Evaluation							Medical Dataset		Open Access	
References	Year	Model	Task	DWS	LDS	JDS	IDRS	PP	DU	Qual	Dataset	Dataset size (N/R)	Dataset	Code
Generation of Diverse types of EHRs														
[37]	2017	medGAN	Generating discrete tabular EHR data	✓	✗	✗	✓	✓	✗	✓	MIMIC-III Sutter PAMF Sutter Heart failure Cohort	46,520/ NA 258,559/ NA 30,738/ NA	✓* ✗ ✗	✓
[57]	2017	RGAN, RCGAN	Generating continuous time-series EHRs	✗	✗	✓	✗	✓	✓	✓	Philips eICU	17,693/ NA	✓*	✓
[180]	2017	GAN for DLEs	Generating continuous time-series Drug Laboratory Effects (DLEs)	✗	✗	✓	✗	✗	✗	✓	Private New York EHR dataset	4,830/ NA	✗	✗
[193]	2018	RadialGAN	Leveraging multiple tabular datasets by using multiple GAN	✗	✗	✗	✗	✗	✓	✗	14 RCTs from MAGGIC	528-13279/ NA	✓*	✓
[10]	2019	medWGAN, medBGAN	Itegrating medGAN with WGAN-GP & BGAN for generating discrete tabular EHRs	✓	✗	✗	✓	✗	✗	✗	MIMIC-III NHIRD Taiwan	46,517 / 46,517 498,909 / 498,909	✓* ✗	✓
[34]	2019	WGAN	Generating heterogeneous discrete tabular EHRs	✓	✗	✓	✗	✓	✓	✓	NMDS	NA / 2,873,466	✗	✗
[172]	2019	SC-GAN	Generating continuous sequentially coupled time-series EHRs data for patient sate & medication dosage	✓	✗	✗	✓	✗	✓	✓	MIMIC-III	29,278 / NA	✓*	✗
[204]	2020	EMR-WGAN, EMR-CWGAN	Improved EHR generation training stability and evaluation	✓	✓	✓	✓	✓	✗	✗	VUMC Synthetic Derivative	2,246,444 / NA	✗	✗
[67]	2020	MC-medGAN	Generating multi-categorical tabular EHRs	✓	✓	✗	✓	✓	✗	✓	SEER's research dataset	NA / 366,631	✓*	✓
[183]	2020	HGAN	Generating Heterogeneous tabular EHRs while preserving feature constraints and inter-dimensional dependencies	✓	✓	✓	✓	✓	✗	✓	VUMC Synthetic Derivative	928,089 / NA	✗	✗
[185]	2019	GcGAN	Generating tabular EHRs while preserving grouped coorlations	✓	✗	✗	✓	✗	✓	✗	Private Pediatric EHR	NA /17,000	✗	✗
[165]	2020	CorGAN	Correlation-Capturing generation of continuous and discrete tabular EHRs	✓	✗	✗	✓	✓	✓	✗	MIMIC-III UCI Epileptic Seizure Recognition	NA / 46,000 500 / 11,500	✓* ✓	✓
[142]	2020	SmoothGAN	Generating tabular EHRs with smooth conditions	✓	✗	✗	✓	✓	✓	✓	Cerner HealthFacts	NA /47,412	✗	✗
[203]	2021	SynTEG	Generating discrete time-series EHRs of diagnostic events	✓	✓	✗	✗	✓	✓	✗	VUMC Synthetic Derivative	NA / 2,187,629	✗	✓
[111]	2020	DAEE	Generating time-series EHRs of discrete diagnostic codes	✗	✗	✓	✗	✓	✗	✓	MIMIC-III UT-Physicians	7,537 / 19,993 13,025 / 85,845	✓* ✗	✓
[112]	2021	EHR-M-GAN	Generating mixed-type time-series EHRs	✓	✗	✓	✓	✓	✓	✓	MIMIC-III Philips eICU HiRID	28,344 / 28,344 99,015 / 99,015 14,129 / 14,129	✓* ✓* ✓*	✓
[174]	2021	WGAN	Federated learning for GAN for discrete, binary EHRs	✓	✗	✓	✗	✓	✗	✓	MIMIC-III	NA / 46,520	✓*	✗
Semi-Supervised Learning and Data Augmentation														
[30]	2017	ehrGAN,SSL-GANs	Augmenting data for imbalanced SSL tasks using EHRs of sequences of diagnosis codes	✗	✗	✗	✓	✗	✓	✓	Private insurance dataset	218,680 / 14,969,489	✗	✗
[113]	2018	GAN for SSL	SSL based GANs for detecting rare diseases in unlabelled tabular discrete & continuous EHRs	✗	✗	✗	✗	✗	✓	✗	IQVIA Rx & Dx	2,961,750 / NA	✗	✓
[196]	2019	GAN for SSL	SSL based GANs for detecting rare diseases in unlabelled time-series EHRs	✗	✗	✗	✗	✗	✓	✗	IQVIA Rx & Dx	1,792,760 / NA	✗	✗
[187]	2019	GAN	SSL based labeling of unlabled data, and GAN-based data augmentation in tabular EHRs	✗	✗	✗	✗	✗	✓	✓	20 datasets from UCI Cerebral stroke private dataset	NA / 80-2,000 11,039 / NA	✓ ✗	✗
Imputation of Missingness														
[192]	2018	GAIN	GAN-based discrete & categorical tabular data imputation	✗	✗	✓	✗	✗	✓	✗	UCI Breast dataset	NA / 569	✓	✓
[200]	2018	Stackelberg GAN	Stabilizing GAIN imputation for discrete, continuous, & categorical EHRs using Stackelberg principles	✗	✗	✗	✗	✗	✓	✗	MIMIC-III	38,645 / 58,000	✓*	✗
[116]	2018	GAN with GRUI	GAN-based multivariate time-series EHR imputation	✗	✗	✗	✗	✗	✓	✗	PhysioNet Challenge 2012	NA/ 4,000	✓	✓
[117]	2019	E^{2}GAN	Improved GAN-based multivariate time-series EHR imputation	✗	✗	✗	✗	✗	✓	✗	PhysioNet Challenge 2012 dataset	NA/ 4,000	✓	✓
[188]	2019	Categorical GAIN	Improving GAIN imputation of categorical tabular EHRs	✗	✗	✗	✗	✗	✓	✗	UCI Breast Cancer PRAEGNANT study	NA/ 286 1234 / NA	✓ ✗	✗
[24]	2019	GAIN adaptation	Improving GAIN imputation of mixed tabular EHRs, including multi-categorical features	✗	✗	✗	✗	✗	✓	✗	UCI breast dataset	NA / 569	✓	✓
[46]	2021	MI-GAN	GAN-based multiple imputation for categorical time-series EHRs	✗	✗	✗	✗	✗	✓	✗	ADNI dataset	NA / 649	✓*	✗
[74]	2021	Bi-GAN	Concurrent imputation and prediction in time-series EHRs	✗	✗	✗	✗	✗	✓	✗	Nemours Pediatric All of Us	66,878 / NA 34,226 / NA	✓* ✓	✓
Treatment Effect Estimation														
[194]	2018	GANITE	Generating missing counterfactual data and individualized treatment effects estimation in tabular EHRs	✗	✗	✗	✗	✗	✓	✗	Twins IHDP	11,400 / 11,400 747 / 747	✓ ✓	✓
[125]	2018	CWR-GAN	Generating time-series post-treatment outcomes for ITE estimation in biomedical translation tasks	✗	✗	✗	✗	✗	✓	✗	MIMIC-III	2,000 / NA	✓*	✓
[63]	2020	MGANITE	Estimating effects of continuous, binary and categorical treatments via conditional GANs on tabular EHRs	✗	✗	✗	✗	✗	✓	✗	AML dataset	NA/212	✓*	✓
[114]	2020	GAD	Continuous treatment effect estimation by deconfounding in tabular EHRs	✗	✗	✗	✗	✗	✓	✗	Twins	4,821 / NA	✓	✗
[65]	2021	PSSAM-GAN	Propensity score augmentation matching for tabular EHRs	✗	✗	✗	✗	✗	✓	✗	S. aureus dataset IHDP	NA / 2,006 747 / 747	✗ ✓	✓
Privacy Preservation														
[178]	2018	DPGAN	Generating differential private EHR data using moment-accounting techniques	✓	✗	✗	✓	✓	✗	✗	MIMIC-III	NA / 46,520	✓*	✓
[99]	2018	PATE-GAN	Generating differential private tabular data using PATE	✗	✗	✗	✗	✓	✓	✗	UCI Epileptic Seizure Recognition Kaggle Cervical Cancer UNOS Transplant MAGGIC	NA / 11,500 NA / 858 NA / 23,706 NA / 30,389	✓ ✓ ✓* ✓*	✓
[11]	2019	AC-GAN	Generating Differntially private GAN via discriminator clipping for tabular EHRs	✓	✗	✗	✓	✓	✓	✓	SPRINT MIMIC-III	6,502 / NA 8,260 / NA	✓* ✓*	✓
[173]	2020	PART-GAN	Improving private GAN training of time-series EHRs	✓	✗	✓	✗	✓	✗	✗	Philips eICU	200,000 / 224,026,866	✓*	✗
[190]	2020	ADS-GAN	Anonymizing generated tabular EHR data while minimizing patient identifiability	✓	✗	✓	✓	✓	✓	✓	MAGGIC (RCT data) 3 UNOS Transplant datasets	30,389 / NA 23,706-56,822 / NA	✓* ✓*	✓
[182]	2020	HealthGAN	Improved End-to-End privacy-preserving WGAN-GP with a focus on privacy & resemblance metrics	✓	✗	✓	✗	✓	✓	✓	MIMIC-III	NA / 27,000	✓*	✓
[92]	2021	HCGAN	Improving robustness to privacy attacks by training Cramér GANs for tabular EHRs	✗	✗	✗	✗	✓	✓	✗	UCI Breast dataset Texas Hospital Data	NA / 699 NA / 186,976	✓ ✗	✗

[1] The included evaluation components are (DWS): Dimension-wise Similarity, (LDS): Latent Distribution Similarity, (JDS): Joint Distribution Similarity, (IDRS): Inter-dimensional Relationship Similarity, (PP): Privacy Preservation, (DU) Data Utility, and (Qual) Qualitative Evaluation, which are explained in details in Section 5. [2] The dataset size is reported in the format of (N/R) where N refers the number of patients an R refers to number of records, reported in each of the works. [3] The symbol ✓* refers to data sources that can be accessed after going through an application process.

data [22, 67]. To achieve this aim, the authors used a gumbel-softmax activation function to enable backpropagating for random samples of discrete variables, which has notable improvements for multi-categorical features [93].

Other researchers focused on improving the capturing of local correlations in tabular EHRs by proposed **Correlation Capturing GAN (CorGAN)** [165]. *CorGAN* combined Convolutional GAN and Convolutional Autoencoders to capture the local correlation between features in both discrete and continuous data. Another example is **Grouped Correlational GAN (GcGAN)** [185] where the authors introduced encoders inspired by CorrNN [29] to capture the coorelations between the treatment and diseases. Specifically, GcGAN focused on learning a disease vector that embeds the curative effect based on predefined treatment efficacy outcomes. Unlike previous works on tabular EHRs, SMOOTH-GAN (Sharp sMOOTh eHr) [142] focused on generating laboratory values and medications conditioned on diagnosis codes. The architecture of SMOOTH-GAN relied on a conditional generator; however, they introduced smooth labels to allow for more flexibility when generating various disease stages. By using a random forest–based heuristic function that estimates the condition in a continuous space, they better represent the disease stage and the quality of the synthetic data.

More recent works focused on improving the training stability, such as the work proposed in **EMR Wasserstein GAN (EMR-WGAN)**. The authors removed the autoencoder that was inherited from medGAN to account for discrete features, applied a filtering strategy to enhance GANs training for low-prevalence clinical concepts [204]. With the new changes, EMR-WGAN was able to generate high-fidelity data with reduced noise and improved training stability [204].

4.1.2 Generating Time-series EHRs. While it is useful to generate tabular EHR data that presents patients' state at a single timepoint, tabular data do not capture the dynamics and changes effectively compared with time-series data, in which variables are recorded along a series of timepoints. To address this issue, a framework for Synthetic Temporal EHR Generation was recently presented where the authors focused on generating timestamped diagnostic events (ICD-codes) [203]. Their architecture addresses the problem in a two-stage approach. The first stage sequentially extracts temporal patterns from visits and adopts a self-attention layer [169]. The second stage generates data conditioned on the learned patterns using Wasserstein GAN [5]. In a similar application, the authors proposed to synthesize sequences of EHRs from patients' chronological visits by using the **dual adversarial autoencoder (DAAE)** along with two GANs components [111]. By utilizing the recurrent autoencoder-based generators, DAAE can synthesize sequences of set-valued medical records such as diagnosis ICD-codes. Another GAN adaptation for continuous time-series EHRs was that of Reference [180], whose work generated time-series **drug laboratory effects (DLEs)** trajectories. Their work has many applications for monitoring patients after exposure to interventions, which can prevent adverse drug reactions [161]. In Reference [57], the authors worked on a model to generate continuous time-series EHR data using **Recurrent GANs (RGAN)**, and its conditional generative version, RCGAN. A kind of recurrent neural network, **Long Short-Term Memory (LSTM)** [86], was used for both the generator and discriminator of RCGAN, which are commonly used models for sequential data tasks [86, 121]. Motivated by the clinical practice of dosage adjustments based on patient state and that both components have a mutual influence on each other, Reference [172] developed **Sequentially Coupled GAN (SC-GAN)**. Their model has two distinct LSTM-based generators that coordinate the generation of patient state and medication dosage data. The output of the patient-state generator is fed to the dosage generator, which mimics the clinical practice of assigning dosage based on the patient status [172].

4.1.3 Generating Heterogeneous EHRs. To mimic the heterogeneous nature of EHRs, which include various types (including demographic information, ICD codes, vital sign time series, etc.), developing GANs that target synthesizing mixed-type EHRs and capturing the dependencies between various features is of vital importance. In Reference [34], the authors used WGAN to generate discrete tabular EHR data containing both administrative and diagnostic data, which they referred to as heterogeneous EHRs.

In parallel, Reference [183] developed a model to account for constraints and preserve relationships across generated heterogeneous tabular EHRs that combined binary, categorical, and continuous values. To do so, the authors incorporated penalization for the violation during GAN training [183]. To simultaneously generate continuous-valued and discrete-valued time-series EHRs, GANs for synthesizing Mixed-type longitudinal EHR data (EHR-M-GAN) was lately proposed [112]. The authors utilized a dual variational autoencoder to generate a shared latent space representation of mixed EHR types. In addition, a sequentially coupled generator implemented with bilateral LSTM was adopted during data generation to capture the temporal correlations between heterogeneous types of EHRs.

4.2 Semi-supervised Learning and Data Augmentation

It is often the case in healthcare datasets that different outcome classes are not equally prevalent, as seen in mortality and rare disease prevalence; this issue is referred to as class imbalance in the machine learning domain [94]. Another commonly seen issue is the absence of labels for some samples, which is referred to as unlabelled samples. Learning from both labelled and unlabelled data gained increasing attention in the machine learning community, where **semi-supervised learning (SSL)** approaches such as classification and clustering have proven to be effective in various applications [206].

Some researchers extended the GANs' role for SSL problems by forcing the GANs to output class labels for unlabelled samples [137, 148]. In their proposed setup, a GAN-based model is trained on a dataset with the samples belonging to one of K classes, where there's a high percentage of unlabelled samples. Then the discriminator's role is adjusted to predict which of K+1 classes the samples belong to, where an extra class refers to the synthetic samples [137, 148]. The extension of the discriminator's role to predict classes opened the door for many applications with a high prevalence of unlabelled samples, such as the case of rare diseases where misdiagnosed or delayed diagnosis is common [107]. The work proposed in Reference [113] extended the discriminator's goal to finding the class assignment of real EHR samples to be able to detect rare diseases in a majority unlabelled tabular dataset. In addition, the authors used a modified loss for their generator, where the objective is to generate samples with minimal divergence from the target distribution. This objective is achieved by over-representing samples with low densities in the original distribution, referred to as "complement samples" as initially proposed in Reference [47]. Based on the success of Reference [113], the authors extended the work of GANs for SSL for predicting rare diseases to be compatible with longitudinal data [196]. The main modification to the GAN models was the usage of RNNs for both the generator and discriminator architecture, which allowed for time-series generation.

GAN-based data augmentation methods have been proposed to mitigate imbalanced and unlabelled data. In such cases, generated data from a specific class can be used in conjunction with the real data to improve model performance, generalizability and decrease over-fitting [28]. Data augmentation can be beneficial when the target dataset has highly unlabelled points or is severely imbalanced, as seen in semi-supervised learning applications. For instance, Reference [30], modified the original GANs and proposed *ehrGAN* to learn the transition distribution of the samples by using a generator with variational contrastive divergence [197]. ehrGAN is then used as a part of the loss function of a semi-supervised learning GANs framework SSL-GAN to augment the training data in a semi-supervised learning manner for sequences of diagnosis codes. By learning the transition distribution of real samples, rich structures of the data manifold around true examples are utilized in SSL-GAN to improve performance.

In a similar application, Reference [187] simultaneously addressed the problems of both the unlabelled and unbalanced data by using a GAN-based approach. The authors presented a framework in which the GAN takes labelled data as inputs and uses it to generate new samples. The generated labelled data are then used to train two independent classifiers to predict sample labels. Next, the authors used the classifiers' predictions to assign pseudo-labels for unlabelled samples. Samples with the same pseudo-label predictions from both classifiers are added to the labelled set. The authors then use GANs again to generate new samples in an attempt to re-balance the minority class labels [187]. The final augmented dataset was used to train a classifier that achieved superior performance on various benchmark datasets. It is worth noting that in many of the semi-supervised uses of GANs, the generated data distribution does not need to match the real data distribution, since the objective might be to over-represent the minority class [47].

4.3 Imputation of Missingness

Handling missing data remains one of the major challenges when dealing with EHRs, where data can be highly missing for various reasons. Using incomplete data for training machine learning algorithms can harm their performance, especially in cases where algorithms may not be robust to missingness [122]. Missing data are usually regarded as one of the following depending on the missingness pattern, **missing completely at random (MCAR)**, **missing at random (MAR)**, and not missing at random [115]. In the healthcare domain, missingness can come in any of the three types, depending on the underlying missingness cause. Examples of healthcare-related causes of missing data in EHRs include data recording errors and machine failure, irregular sampling and inconsistent medical visits [108], unmeasured lab tests due to the lack of medical need [54], or even high cost and dangerous to acquire information such as invasive or radiology procedures [20, 104] and other factors related to patient severity and diagnosis [2].

GANs are naturally suitable for generative tasks not only generating completely new samples but also generating missing values that can be used to impute the original samples. While most data imputation methodologies are often based on either parametric or non-parametric probability density estimation, GANs can perform data imputation without calculating a probability density first [188]. The first GAN-based missing data imputation paper had a focus on image completion [90]. This work motivated a series of GAN-based imputation methods that are application-specific and tailored for various data types including medical data. For instance, Reference [192] proposed the use of an adjusted version of the original GANs that they refer to as **Generative Adversarial Imputation Nets (GAIN)**. In their work, the generator's role was adjusted to generate and accurately impute missing data. The discriminator's role, however, was adjusted to distinguish between original and imputed components, analogous to distinguishing between real and synthetic samples [192]. To increase the performance and the quality of the generated imputation data, the discriminator is also given additional information "hints," which reveals to the discriminator partial information about the missingness of the original sample. Their work focused on MCAR missingness in multiple tabular datasets. The results of GAIN were benchmarked against various data imputation methods such as MICE [168], missForest [158], and Expectation-maximization [134].

Others were motivated by the high missingness in the commonly used EHR data such as the *MIMIC III* dataset [97], where missingness reached as high as 74% [200]. Reference [200] combined the structure proposed in Reference [192] with principles of Stackelberg competition in the domain of game theory [60]. The main adaptations of GAIN are in the use of multiple generators (followers), rather than one, which team up against the discriminator (leader). Their results showed that the Stackelberg-GAN was able to capture complex data distributions and achieved high performance when compared with other state-of-the-art imputation methodologies. The authors evaluated their work on discrete, continuous, and categorical tabular EHRs [200].

In a similar work, Reference [188] proposed a modification to GAIN that focused on improving performance in generating categorical tabular EHR data. The authors hypothesized that the original GANs architecture and the one used in Reference [192] is not optimal for categorical features due to the softmax function's ability to produce values between 0 and 1 [188]. To address this, Reference [188] introduced a fuzzy binary coding of categorical features, where values are encoded using real numbers between 0 and 1 to preserve the categorical information aspect of the data. To further improve GAIN for mixed-type tabular EHRs, Reference [24] modified its model structure where the generator and discriminators had multiple inputs as well as multiple outputs [24]. The major contributions focused on variable splitting and the usage of gumbel-softmax activation that accounts for categorical variables and their discrete distributions [93]. While most works focused on MCAR cases, the authors of **Multiple Imputation via GANs (MI-GAN)** introduced an architecture that is theoretically supported for both MCAR and MAR cases. The authors combined ideas from both GAIN and Multiple Imputation machine learning works to solve the problem of MAR blockwise pattern missingness where the missing probabilities depend on the observed values in the dataset [46]. The results showed superior performance with respect to other imputation models in terms of statistical inference and computational speed.

Despite the outstanding results of GAIN and its various adaptations, they are not directly applicable to time-series EHRs. To fill this gap, the authors of Reference [116] proposed a GAN-based model that is implemented with a modified **Gate Recurrent Unit (GRU)** [35] to model the temporal irregularity of the incomplete time-series data, which they refer to as **GRU for data Imputation (GRUI)** cell. The use of GRUI, instead of LSTM and other RNN variants is motivated by its compatibility with the irregular time lags and variations between two consecutive observations including those seen in data such as ICU EHRs [116]. GAN with GRUI model performs the imputation in a two-stage approach. First, it trains a GAN model to generate complete time series, and then it tries for each sample to find the "noise" vector that is most similar to the original sample [116]. Despite reporting state-of-the-art results for imputing time-series EHR data, the work of Reference [116] has a major drawback in terms of training efficiency. Motivated by improving the efficiency of GAN for time-series imputation, Reference [117] proposed an end-to-end GAN-based imputation model, referred to as E^2GAN. The proposed model performed imputations with reduced training time, with higher quality by adopting a compressing and reconstructing strategy to circumvent the noise optimization stage in the GAN with GRUI [116]. Recently, Reference [74] presented a novel GANs architecture *Bi-GAN* to perform both imputations of missing values and prediction of future values in time-series EHR data. Both the generator and discriminator were bi-directional recurrent neural networks Bi-RNNs, which are suitable for time-series applications. In their work, the GAN-based model learns from all the observed samples to impute missing values and then learns to predict future values by treating them as missing values [74]. This problem setup does not require a definition of prediction windows at training time, which motivates flexible predictive models that they refer to as "any-time prediction tool" [74].

4.4 Treatment Effect Estimation

Estimating treatment effects is a complicated causal inference task with many data challenges, where the aim is to estimate the patient's response to a specific treatment. The major challenges in this field arise from missing counterfactual data, the unobserved outcomes of untaken treatments [83]. In **Randomized Control Trial (RCT)** settings, patients in the treatment group are matched to those in the control group to compensate for missing counterfactuals. However, despite being the golden standard for various clinical applications, RCT-based treatment effect estimation suffers from multiple issues concerning their high cost [150], relatively small size [77], ethical issues [66], and short duration of followups that might miss out long-term effects of medications [17]. A low-cost alternative to RCT data is the regularly collected EHR data. Specifically, longitudinal EHRs, which include diverse patient cohorts, long-term outcomes with no strict exclusion criteria, making EHRs more representative of the patient population [17, 135]. However, in EHR data, treatments are not assigned at random, and there is no clearly defined control group. Thus, estimating treatment effects from EHRs requires measures to control confounding effects and perform covariate adjustment [126, 144] to avoid selection bias.

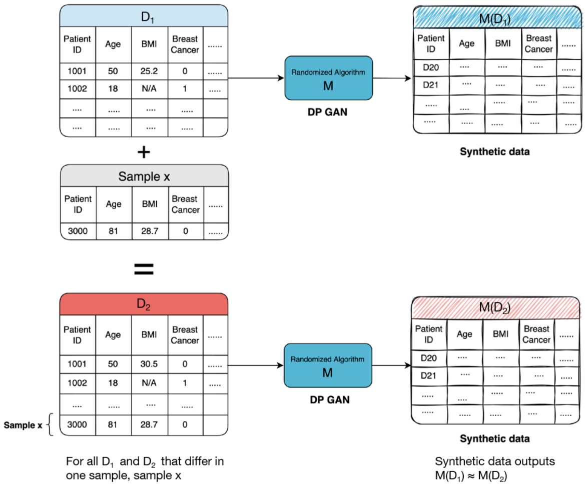
The generative capabilities of GANs are a valuable option for various treatment effect estimation applications. In Reference [194], the authors made use of GANs' generative properties to generate counterfactual outcomes. In their novel design, **GANs for inference of Individualized Treatment Effects (GANITE)**, they considered counterfactual outcomes to be missing labels, similarly to their earlier work in Reference [192]. GANITE utilized a pair of GANs: one for counterfactual imputation and another for treatment effect estimation. In the first GAN, the generator's task is adjusted to generate missing counterfactual outcomes, while the discriminator's task is to tell the factual from the counterfactual outcomes. In the presence of counterfactual outcomes, a treatment effect estimation function can be predicted using traditional machine learning models. However, in GANITE, the authors utilize another GAN to model treatment effect estimation by taking the output of the counterfactual GAN as input and generating a potential outcome vector with confidence intervals [194]. While GANITE focused on binary treatment, Reference [40] focused on generating time-series post-treatment outcomes. The authors' work was motivated by the scarcity of paired pre- and post-treatment patient time-series data in settings such as ICU ventilation and vasopressors assignment. Their proposed model, **Cycle Wasserstein Regression GAN (CWR-GAN)**, is a hybrid of several architectures: original GAN [69], Wasserstein GAN [5], and cycle-consistent GAN [205]. The authors of CWR-GAN tested their model in regression-based tasks and provided an alternative to the traditional uni-directional regression approaches, where unpaired data would be ignored during training [40].

To extend GAN-based treatment estimations from binary to various kinds of treatment variables including categorical and continuous, Reference [63] applied modifications to GANITE, which they named *MGANITE*. Estimating continuous treatment is of high importance in applications involving dosage adjustment especially in oncology [127]. One of the main modifications was a mathematical adjustment to the loss function that takes a treatment assignment vector in both the counterfactual and ITE estimation blocks to allow for simultaneous treatment effect estimation [63]. When using observational data where treatments are not randomly assigned, controlling the confounding factors, such as using propensity scores is essential [45]. In Reference [114], the authors propose a GAN-based model that generates a "calibration" distribution, one that eliminates associations between covariates and treatment assignment by a random perturbation process of the treatment variable. The generative capabilities of GANs are used to learn a weight vector that is used to adjust the distribution of observed data and construct the calibration data. The authors refer to their model as **Generative Adversarial De-confounding (GAD)** [114].

Statistical approaches such as **propensity score matching (PSM)** are commonly used by classical treatment effect estimation works to balance the population's characteristics assigned either to an intervention or a control group [21]. However, despite their popularity, PSM approaches can lead to high reductions in sample sizes due to unmatched control samples [21]. Lately, a GAN-based propensity score synthetic augmentation matching model, *PSSAM-GAN*, was proposed to mitigate the problem of sample size reduction using PSM approaches [65]. First, the authors matched their samples based on calculated propensity scores. Then, to be able to use unmatched samples, the authors used a GAN-based model to generate treatment matches for the unmatched control samples [65]. Finally, the original EHR data were augmented with the newly generated matched samples to be used for downstream treatment estimation tasks [65].

4.5 Privacy Preservation

Privacy is a central theme in GAN development, as it is a principal motivator for using generative models in healthcare applications. Even though GANs do not explicitly expose patient data, some works demonstrated the importance of improving the privacy preservation of GANs, especially when dealing with sensitive information such as patient EHRs [133]. In the field of privacy, there



D1

Patient ID	Age	BMI	Breast Cancer	......
1001	50	25.2	0	......
1002	18	N/A	1	.....
....	.....	....	.....	
....	.....	....	.....	

Sample x

Patient ID	Age	BMI	Breast Cancer	......
3000	81	28.7	0	......

D2

Patient ID	Age	BMI	Breast Cancer	......
1001	50	30.5	0	......
1002	18	N/A	1	.....
....	.....	....	.....	
3000	81	28.7	0	

M(D1)

Patient ID	Age	BMI	Breast Cancer	......
D20	....	....	....	......
D21	....	....	....	......
.....	....	.....	....	.....
.....	.....	.....	....	.....

M(D2)

Patient ID	Age	BMI	Breast Cancer	......
D20	....	....	....	......
D21	....	....	....	......
.....	....	.....	....	.....
.....	.....	.....	....	.....

Fig. 3. GAN training with differential privacy guarantees. Real datasets D_1 and D_2 only differ in a single sample X. M is the differentially private GAN model that outputs $M(D_1)$, and $M(D_2)$, which at most have a difference of $\mathbf{e}^{\epsilon}$.

has been a wave of frameworks that apply theoretical guarantees to ensure the privacy of the data [166]. Notably, differential privacy is a theoretical guarantee that allows learning nothing about an individual while learning useful information about a population [55]. Differential privacy is concerned with the impact of the presence or absence of a single record on the outcome of the computational tasks. Differential privacy is defined as follows.

A randomized algorithm M is ϵ-differentially private if for any two datasets D_1 and D_2 that differ in a single point and for any subset of outputs S,

$$P(M(D_1) \in S) \leq e^{\epsilon} P(M(D_2) \in S),$$

where P is taken with respect to the randomness, $M(D_1)$ and $M(D_2)$ are the outputs of the M for databases D_1 and D_2, respectively [55]. Based on this definition, there are many differentially private algorithms, any of which may be used to complete the same computational task under differential privacy guarantees [55]. Differential privacy can be applied to GAN training, where M refers to the differntially private GANs as seen in Figure 3.

Motivated by improving privacy through providing theoretical guarantees for medical data, several works developed and evaluated differentially private GANs for EHRs generation applications. Namely, DPGAN [178] proposed GANs with differentially private guarantees by adding noise to the discriminator's gradients, which was inspired by moment accountant techniques [1]. Similarly, Reference [99] proposed a modification to the GAN training of the discriminator by using an adaptation of the differentially private framework **Private Aggregation of Teacher Ensembles (PATE)** [138]. In PATE, multiple teacher models are independently trained on subsets of the data for a classification task. The final classification output is an aggregate of each of the teacher model's prediction [138]. Another differentially private GANs for EHRs development was Reference [11], where the authors limited the effect of a single participant on the training by clipping the norm of the discriminator's gradient combined with the addition of Gaussian noise. In a similar spirit, the authors of Reference [173] proposed a data augmentation framework with differential private guarantees and model optimizations to improve the data utility without compromising the quality. The proposed framework, **privacy-preserving Augmentation and Releasing**

scheme for time-series data via GAN (PART-GAN), uses weight pruning and grouping, generator selecting, and denoising mechanisms for improving the quality in time-series data [173]. Some works combined both theoretical and empirical evaluations to prove the privacy preservation of the GAN model [92]. To avoid compromising the synthetic data fidelity, the authors applied partial differential privacy to the Quasi Identifier features; these features are then recombined with the other sensitive attributes. The authors then trained a GAN that relies on Cramér distance [12] between the joint distribution of the generated observation and real differentially private patient data using the combined feature set. The model was then tested for various adversarial attacks to support their theoretical guarantees [12].

Despite the strong privacy guarantees of differential privacy, it has various technical limitations such as compromised data fidelity and utility. This motivated works to look for strong privacy-preserving alternatives. For example, Reference [190] developed a Wasserstein GAN with a Gradient Penalty-based [72] model that they refer to as **anonymization through data synthesis using generative adversarial networks (ADS-GAN)**. In their work, the authors created a mathematical definition for "identifiability," which was based on the probability of re-identification given the combination of all data on any individual patient [190]. In *ADS-GAN*, the authors tested for the data quality, while maintaining the identifiability constraints. In a similar notion, the authors of HealthGAN [182] worked on an end-to-end privacy-preserving GAN based on WGAN-PB and proposed a quantitative privacy metric, $\text{privacy}_{\text{loss}}$ that is based on the balanced accuracy of the adversarial nearest-neighbors model. The work of HealthGAN was later extended and evaluated in various settings [48, 181].

5 EVALUATION OF GANS FOR EHRS

Despite the substantial attention given to theoretical and application-oriented GAN development gained over the past years, there is still no consensus on evaluation metrics or methodologies [163]. Evaluating the strengths and shortcomings of the model and synthetic data is vital for fair benchmarking and future research directions. For example, evaluating whether the GAN model is simply memorizing training examples or is missing important information and characteristics relating to data distribution is essential prior to using the synthetic data for downstream tasks. The evaluation of GAN models can take various directions all of which have different aims such as close approximation of data distribution, maintaining privacy, the utility for downstream machine learning tasks, and model performance. Evaluation methods described in the literature, including those seen in the papers presented in Table 1, can be grouped into two groups: (1) qualitative and (2) quantitative evaluation methods. Previous work by Hernandes et al. [80] grouped evaluation metrics into three categories: resemblance evaluation, utility evaluation, and privacy evaluation. El Emam et al. [56] focused on utility metrics exclusively. In this work, we review the various utility and privacy evaluation metrics; however, we choose to expand on the various types of resemblance/similarity metrics. For this purpose, we categorize the quantitative similarity metrics into four main categories, namely dimension-wise, joint-distribution, inter-dimensional, and latent space similarity, respectively. Dimension-wise similarity refers to the metrics used to measure the similarity between each dimension/feature in the synthetic and real datasets. Joint-distribution similarity, however, measures the distribution similarity across all features and samples. Inter-dimensional distribution similarity focuses on preserving the inter-dimensional relationships in the synthetic dataset. Last, latent space similarity compares the synthetic and real datasets via latent space representation. Other presented quantitative metrics in this work are related to the data utility and privacy preservation of the generated synthetic data. In Table 2, we present a list of the different quantitative evaluation metrics and tests used in the reviewed papers along with the data types each metric was used to evaluate and the corresponding reference of each metric

Table 2. Quantitative Metrics and Tests Used for Evaluating GANs for EHR Models

Reference	**Evaluation metric**	**Data Type**	
		Tabular	Time Series
Dimension-wise Distribution Similarity			
[37]	Dimension-wise Probability	✓	✓
[10]	Dimension-wise Average	✓	NA
[123]	Kolmogorov–Smirnov (K-S) test	✓	NA
[67]	Support Coverage	✓	NA
[13]	Kullback–Leibler Divergence (KLD)	✓	NA
Joint Distribution Similarity			
[13]	Kullback–Leibler Divergence (KLD)	✓	NA
[119]	Jensen–Shannon Divergence (JSD)	✓	NA
[146, 167]	Wasserstein Distance	✓	NA
[70]	Maximum Mean Discrepancies (MMD)	NA	✓
[148]	Inception Score (IS)	NA	✓
[183]	Cross-type Conditional Distribution	✓	NA
[204]	First-Order Proximity	✓	NA
[191, 202]	Discriminative Score	NA	✓
Inter-dimensional Relationship Similarity			
[37]	Dimension-wise Prediction	✓	NA
[14]	Pairwise Pearson Correlation	✓	✓
[198]	Association Rule Mining (ARM)	✓	NA
[183]	Frequent Association Rules (FAR)	✓	NA
Latent Distribution Similarity			
[204]	Latent Space Representation	✓	NA
[203]	Weighted Latent Difference	NA	✓
[176]	Log-cluster	✓	NA
Data Utility			
[57]	Train on Synthetic, Test on Real (TSTR)	✓	✓
[57]	Train on Real, Test on Synthetic (TRTS)	✓	✓
[111, 203]	Predictive Modeling, Forecast Analysis	NA	✓
[98, 99]	Synthetic Ranking Agreement (SRA)	✓	NA
[172]	Data Augmentation Test	NA	✓
Privacy Preservation			
[55]	Differential Privacy Guarantees	✓	✓
[156]	Member Inference Attack	✓	✓
[124]	Attribute Disclosure Attack	✓	✓
[62]	Model Inversion Attack	NA	✓
[190]	Identifiability	✓	NA
[182]	Privacy$_{\text{loss}}$	✓	NA
[174]	Exact-matches Test	✓	NA

[1] The works referenced in the first column of Table 2 refer to the papers that explain the respective evaluation metrics. [2] The ✓symbol refers to metrics that were utilized to evaluate synthetic data for the corresponding data type, while NA refers to those with no available validation in the literature.

5.1 Quantitative Evaluation

5.1.1 Dimension-wise Distribution Similarity. A major objective of generative models is generating data whose distribution highly resembles that of the real dataset. Many evaluation metrics have been proposed to quantitatively evaluate the distribution resemblance per feature or "dimension." For instance, dimension-wise probability is a test that compares the probability distribution of each of the features in real and synthetic datasets. The comparison method varies depending on the structure and type of data. For example, the Bernoulli success probability or Pearson Chi-square test were used for binary features [37, 178, 183, 190, 204], while in other works the Student t-test was used continuous variables [190]. A similar evaluation test, Dimension-wise Average, was introduced to account for discrete count variables such as disease or procedure codes. The test simply calculates the dimension-wise average and compares that of the real to the synthetic dataset [10]. Another commonly used test is the Kolmogorov–Smirnov K-S test, which simply tests that two samples came from the same distribution [123]. The test is based on a well-known statistical metric, which is calculated by finding the maximum absolute value of the differences in the cumulative distribution functions of the two compared samples as seen in Reference [10]. Other works took less rigorous evaluation approaches by reporting the distributions and statistical values as mean and standard deviation of both the synthetic and real datasets [11, 173]. To measure the extent of variable distribution coverage in the synthetic data, Reference [67] used support coverage metric. In this metric, the ratio of the cardinalities of a variable's support is calculated in the real and synthetic data. The final result aggregates the results of the over all variables to measure the joint support coverage. While more commonly used to measure overall data divergence, some papers used divergence metrics such as the **Kullback–Leibler Divergence (KLD)**, which is also known as the relative entropy on the feature level, as seen in Reference [67]. KLD is used for many applications to calculate a score or distance that quantifies the divergence of one probability distribution from another [13]. KLD is seen for many applications including Gaussian Mixture Models and t-distributed stochastic neighbor embedding. By definition, KLD is defined to be

$$\boldsymbol{KLD}(\mathcal{P}, \mathcal{Q}) = \int_{\mathcal{X}} \mathcal{P}(x) \log \frac{\mathcal{P}(x)}{\mathcal{Q}(x)} dx$$

for distributions $\mathcal{P}$ and $\mathcal{Q}$ [13].

5.1.2 Joint Distribution Similarity. Preserving the real data distribution is a major aspect of evaluating the GAN quality. Aside from evaluating the distribution at the individual feature level, synthetic data needs to be evaluated in terms of preserving the joint distribution of the real data. Joint distribution is usually evaluated by calculating one of the distance metrics such as KLD [13] as seen in Reference [63]. However, one of the major drawbacks of KLD is that it is not symmetrical, where KLD($\mathcal{P}, \mathcal{Q}$) $\neq$ KLD($\mathcal{Q}, \mathcal{P}$). To overcome this, GAN-based models can be more accurately evaluated using **Jensen–Shannon Divergence (JSD)** [119]. The definition of JSD builds on KLD, which is defined as follows:

$$\boldsymbol{JSD}(\mathcal{P}, \mathcal{Q}) = \frac{1}{2} KLD(\mathcal{P}, \mathcal{M}) + \frac{1}{2} KLD(\mathcal{Q}, \mathcal{M}),$$

where $\mathcal{M}$ is the average distribution with density $1/2 * (\mathcal{P} + \mathcal{Q})$, for distributions $\mathcal{P}$ and $\mathcal{Q}$ [136].

Results using JSD are symmetrized and smooth, which explains its usage in training critic of many GAN applications including the original GAN architecture [69] as well as in the evaluation of some of the GANs for EHR applications [190]. Another KLD-based metric is the **inception score (IS)**, which was introduced in Reference [148], and is commonly used in many imaging applications. Despite capturing the quality and diversity of the data, IS is highly sensitive to noise, and thus it is rarely used in evaluating GANs for EHRs models [173].

Another joint distribution metric used is based on the **Wasserstein distance (WD)**, also referred to as Earth Mover's Distance, which informally measures the minimum mass displacement to transform one distribution into the other [167]. Even though this is a metric used for evaluating the joint-distribution similarity in the synthetic data, it is more often used in training loss function as seen in the well-known Wasserstein GAN, which was introduced to overcome overfitting and mode collapse issues in GAN training [4, 5]. The WD for P and Q distributions over $\mathcal{X}$ is defined as

$$\boldsymbol{WD}(P, Q) = \inf_{\gamma \in \Gamma} \int_{\mathcal{X} \times \mathcal{X}} \|x - y\|_2 d\gamma(x, y),$$

where Γ is the set of all possible joints on $\mathcal{X} \times \mathcal{X}$ that have marginals P and Q [167].

The usage of WD as a training critic has been particularly seen in many GANs for EHRs works reviewed in this article [10, 174, 190, 204], while fewer works used it as an evaluation metric for joint similarity [190]. One major drawback of the WD is that it tends to be intractable in high dimensions, as well as its high computational complexity and biased sample gradients [7, 12, 100]. Another commonly used quantitative evaluation metric is **Maximum Mean Discrepancies (MMD)**, which was first introduced in 2012 as a kernel two-sample test [70]. MMD measures the dissimilarity between two probability distributions and uses samples drawn independently from each distribution [70]. MMD relies on the idea of representing distances between the compared distributions as differences of feature embeddings, mapped using **Reproducing Kernel Hilbert Space (RKHS)** [15, 70]. More formally, MMD between two distributions P and Q over $\mathcal{X}$ in in the the RKHS Kernel $\mathcal{H}_k$ is

$$\boldsymbol{MMD}_k^2(P, Q) := \mathbb{E}_{x,x'}\left[k\left(x, x'\right)\right] + \mathbb{E}_{y,y'}\left[k\left(y, y'\right)\right] - 2\mathbb{E}_{x,y}[k(x, y)],$$

where $x, x' \overset{iid}{\sim} P$ and $y, y' \overset{iid}{\sim} Q$ [70].

Some works proposed novel joint distribution similarity tests that focus on the overall preservation of conditional distribution. For example, the Cross-type Conditional Distribution [183] metric evaluates if the synthetic data maintain the distribution of one data type conditioned on another. The conditional distribution is quantified in terms of the mean and standard deviation and then compared between synthetic and real datasets. **Fist-Order Proximity (FOP)** is another metric introduced in Reference [204] measures the similarity of the structural associations between the real and generated datasets. To do so, an undirected graph is generated in which the weight of an edge between categorical features, such as diagnosis codes, corresponds to their co-occurrence frequency in the population. The difference in FOP between the synthetic data and real data is calculated and used as a metric of preserving the associations. Other researchers evaluated joint distribution similarity using unsupervised clustering-based evaluations can be employed as seen in Reference [180]. Similarly, an unsupervised-based evaluation was introduced in Reference [182], where the adversarial accuracy of a clustering model is used to capture resemblance loss of the GAN model, which the authors refer to as Train and Test resemblance losses.

Other than the aforementioned unsupervised-based evaluation, some authors leveraged an additional supervised task to quantify GANs' performance. A binary classifier (a post hoc discriminator) is trained to discriminate between the synthetic samples and the held-out real samples. The performance of the model, the discriminative score, is used to quantify the synthetic data's resemblance to the real data without calculating statistical distances [111, 112].

5.1.3 Inter-dimensional Relationship Similarity. Other than evaluating for the dimension-wise and joint similarities, it is important to also assess the synthetic data's preservation of inter-dimensional relationships and correlation between features. Several works used the Dimension-wise Prediction test introduced in Reference [37]. This test iteratively chooses a feature and assigns it as a label and treats the rest of the features as inputs. Two classifiers are trained, where one is trained on real data and the other is trained on the synthetic data to predict the selected label [37, 165]. The performance for each of the trained classifiers per feature is then compared, and the assumption is that the closer the performance of pairs, the better the quality and inter-dimensional relationship similarity of the synthetic dataset [10, 37, 165, 178, 183, 204]. The trained classifiers are usually logistic regression models [37, 165], but at other times different classifiers such as **support vector machine (SVM)** and random forest were used [10]. Other works conducted inter-dimensional correlation evaluations such as Pearson Coefficient Correlation matrices comparisons for both real and synthetic data [11, 67, 172, 190]. The resulting mean vector and covariance matrices are compared to evaluate the resulting dataset for preserving inter-dimensional correlations and relationships.

Association Rule Mining (ARM) is commonly used in clinical data-mining applications. ARM models are used to identify meaningful patterns rules among clinical concepts [155, 179]. The GANs' ability to preserve the rules identified in the real set was evaluated by using a machine learning ARM model to identify association rules and compare those derived from the real to those of the synthetic [10]. Other authors introduced **Frequent Association rules (FAR)** [183], which utilizes the theoretical bases of ARM. FAR checks for both support and confidence, where supports represent how frequently the condition set appears in the dataset, whereas confidence is an indication of how often a condition rule is true [179]. After applying ARM, the proportion of the association rules that appear in both the real synthetic data are compared and then reported in terms of classification performance metrics such as precision and recall [179].

5.1.4 Latent Distribution Similarity. Building on the intuition that a good GAN model generates synthetic data that captures lower-level relationships even in the latent space, several works evaluated the latent distribution similarity between the real and synthetic datasets. For example, References [183, 204] used a Latent Space Representation test, where real and synthetic samples are projected into the latent space by utilizing a β variational autoencoder [82]. After obtaining the projection in latent space, the dimensional mean of the distribution variance of each of the latent features is calculated in the synthetic data and compared to that of the real counterpart. A smaller distance or difference corresponds to a higher resemblance. This metric becomes of higher relevance when considering applications where interpretability is an integral component. Latent space evaluation metrics were also used in Reference [203], where the authors calculated a weighted K-S average across all latent features. The latent space presentation and weights were arrived at by applying Singular Value Decomposition [106], which yielded singular vectors and the corresponding singular values (weights) for each of the features. The calculated weighted averages for the synthetic and real data were compared to test for similarity in the latent space representation. Another way to measure the similarity in the latent space is by using unsupervised learning approaches such as the log-cluster metric [176] as seen in Reference [67]. To measure the similarity of the underlying latent structure of the real and synthetic data, both datasets are merged and clustered using k-means clustering. Disparities of cluster membership of real samples versus synthetic samples are indicative of latent representation divergence [67].

5.1.5 Data Utility. High-quality synthetic data are a valuable asset for various research purposes, as seen in Section 2. Evaluating the synthetic data in terms of its utility is a practice that has been adopted by many of the works included in Table 1. One of the earlier machine learning utility testing frameworks was proposed in Reference [57], which is **Train on Synthetic Test on Real (TSTR)**. As the name implies, a machine learning model is trained on synthetic data and then tested on held-out real data. Similarly, **Train on Real, Test on Synthetic (TRTS)** was also proposed in Reference [57], as a reverse case of TSTR. When evaluating TSTR, the results show the utility of synthetic data when used for model buildings and conducting analysis; however, the model is applied to real data. However, TRTS could potentially supplement the performance of a model that is trained and tested on real data, with results on a synthetic dataset based on a dataset from a different source, where access to the second dataset might not be feasible. The framework is flexible and can be used on any task-based machine learning application such as supervised classification [57] where classification metrics such as F1 score, accuracy, and precision can be reported on both the synthetic and real datasets [88]. Other works assessed TSTR for supervised regression [30, 34, 203], where metrics such as **Area Under Receiving Operator Curve (AUROC)** and Area Under Precision-Recall Curve [88] were reported. Semi-supervised learning works focusing on mitigating data imbalance issues evaluated the utility of synthetic data for machine learning tasks for the same purpose [30, 113]. Time-series-specific supervised learning evaluations were be applied to generative tasks to evaluate the preservation of temporal dynamics [111, 203]. The same temporal-related supervised task on both the real and synthetic datasets, such as predicting the top-N ICD codes in patient's next visit [111], or forecasting patient's future diagnosis [203], which were referred to as predictive modeling performance or forecast analysis, respectively. A similar performance of the models on both the synthetic and real dataset is indicative of the GANs' ability to preserve characteristics and utility of the real data.

Despite their wide use, TSTR, TRTS and other data utility evaluations are sensitive to the model chosen for evaluation. For example, it may be the case that a logistic regression model performs similarly on both synthetic and real data, but that might not be the case when other models are used, such as SVMs or neural networks. To mitigate this issue, the authors of Reference [99] propose **Synthetic Ranking Agreement (SRA)**, a framework that evaluates a selection of models trained on the synthetic and tested on the synthetic. The performances of the same models are compared to those trained and tested on real data [98, 99]. The authors then define a metric that performs ranking agreement and comparison to evaluate the power of the synthetic data for machine learning downstream tasks. Although this metric can suffer from the same limitation of TSTR and TRTS frameworks, it evaluates a broader range of machine learning classifiers that is a step closer to the ideal machine learning utility assessment.

$$\boldsymbol{SRA}\left(\{A_i\}_{i=1}^{L}, \{C_i\}_{i=1}^{L}\right) = \frac{1}{L(L-1)} \sum_{j=1}^{L} \sum_{k \neq j} \mathbb{I}\left(\left(A_j - A_k\right) \times \left(C_j - C_k\right) > 0\right),$$

where L is a set of predictive models $f_1, f_2, \ldots, f_L^2$. $A_i \in \mathbb{R}$ stands for the performance of the models when trained and tested on the real data, while $C_i \in \mathbb{R}$ stands for the performance of the models trained and tested on the synthetic data [99]. To scale the valuation of the utility of synthetic data for machine learning applications, Reference [182] studied the educational utility by hosting an online challenge for students to evaluate the quality of the data.

It is important to note that some works applied one of the frameworks mentioned here, for example, TSTR, however, they did not explicitly mention the framework's name. In many machine learning applications, synthetic data can be used to augment real data. To evaluate how much synthetic data are needed to achieve the desired performance, Reference [172] presented a Data Augmentation Test, where the authors evaluated the synthetic data's utility for machine learning applications. Similarly, in the performance of models was evaluated using augmented dataset, while varying the percentages of synthetic data used in each variation [112, 185].

Data utility metrics and tests were also employed in non-machine learning tasks to evaluate the synthetic data for its intended utilization. This was specifically seen in GANs for missing data imputation tasks where the GAN-imputed data were evaluated in terms of **Root Mean Square Error (RMSE)** [27] and Mean Absolute Error [27] as shown in Reference [24, 74, 192]. GANs for imputation tasks were also evaluated post-imputation prediction performance in terms of AUROC, FI, and accuracy and benchmarked against other state-of-the-art data imputation techniques as seen in References [74, 188, 192, 200]. Similarly, GANs for estimating treatment effects work were evaluated in terms of Precision in Estimation of Heterogeneous Effect, average treatment effect [83], the average treatment effect on the treated [153], and RMSE for controlling the confounding evaluation [114].

5.1.6 *Privacy Preservation.*

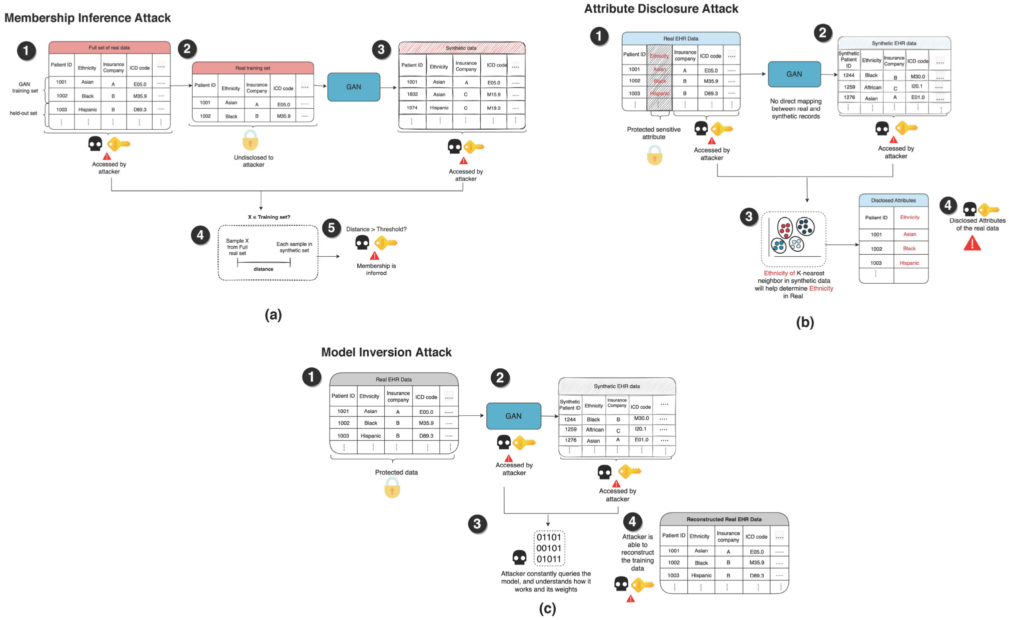
Evaluating the quality and fidelity of the synthetic data is essential. However, to ensure safe usage of the resulting synthetic data, there is also a need to make sure that patients' privacy is not compromised. As there is no universally accepted standard definition for privacy [109], the works included in this article dealt with privacy evaluation in a wide range of ways. Theoretical privacy guarantees such as differential privacy have been used in many of the GANs for EHRs works, as seen in References [11, 34, 57, 99, 173, 178]. With the strict differential privacy's guarantees that neatly confirm privacy preservation, such works generally did not undertake further information leakage evaluation. While such approaches might seem ideal, differential privacy might lead to compromised data and utility preservation [53] as seen in References [34, 57]. An alternative to theoretical guarantees is the empirical evaluation of the robustness to well-studied attacks. The attacks evaluated in the reviewed papers include (a) membership inference attacks, (b) attribute disclosure attacks, and (c) model inversion attacks. First, **membership inference (MI)** attacks assumed that the attacker has access to the records of a set of real patient records and attempts to determine if anyone from the real patients is in the training set of the GAN model [156]. To test for MI scenarios, a distance metric is calculated between each record in both the real and synthetic datasets. A threshold is then chosen as a cutoff, such that any record from the synthetic data with a distance less than the threshold is considered from the training set. Some works calculated this distance using hamming distance [37, 67, 183, 204], while others used cosine similarity [165, 174] and MMD [142]. The performance is then reported in terms of precision and recall to quantify the GANs' robustness to MI attacks. In other instances, a model is used to estimate the likelihood for a given record referred to as perplexity and then metrics such as R^2 and KLD are used to estimate the extent of distribution similarity as a proxy log likelihood [203]. An overview of a sample MI attack is shown in Figure 4(a).

The second type of adversarial scenarios is attribute disclosure attacks that occurs when an attacker can infer additional attributes about a patient by knowing a subset of other attributes about the same patient [124]. To simulate this scenario, a random percentage of the real training set is sampled as well as a random set of features to be those disclosed to the attacker [37]. A voting-based *k*-nearest neighbor clustering classification is utilized to estimate the values of the known features and then performance metrics in terms of precision and recall are reported as seen in References [37, 67, 183, 204]. Some works extended this simulation by assuming the worst-case scenario where the attacker also has prior statistical knowledge about the undisclosed features [203]. An example attribute disclosure attack is shown in Figure 4(b). The other type of attacks, namely model inversion refers to the scenario where an attacker aims to reconstruct the training data by their ability to constantly query the model [62], as shown in Figure 4(c). This kind of attack was not frequently used in GANs for EHRs evaluation [92], due to its replication complexity. The aforementioned attacks can be implemented under two different scenarios against the generative models, either black-box or white-box setting [76]. In a white-box setting, the attacker has full access to the target model, including the architecture and weights of a trained network. While in a black-box setting, the attacker is only able to make queries to the target model and has no knowledge of its internal parameters as implemented in Reference [112].

Some papers also developed a mathematical definition of privacy, such as identifiability, which refers to the probability to re-identify samples included in the training [190]. Similarly, Reference [182] proposed an unsupervised adversarial privacy-based privacy-loss metric to quantify the extent of privacy preservation. Last, simple evaluations such as Exact-Matches test were applied to check for the presence of exact duplicates of the training data in the synthetic data [174].

5.2 Qualitative Evaluation

Qualitative evaluation approaches are commonly utilized in GAN papers to support the quantitative results with reasonable simplistic measures. For example, several papers reported visualization of data distributions and embeddings, such as comparing generated feature distribution plots [182], and correlation heat-maps [30]. While others qualitatively compared patient trajectories by visually comparing the synthetic time-series signals [11, 57, 172]. An example of a qualitative privacy evaluation is the interpolation test proposed Reference [57], where a pair of training are back-projected into the latent space and linearly interpolating them produces smooth variation in the sample space, and then the GAN model is then used to produce samples at each point. The variation in the outputs is used as a proof of the GANs' ability to capture the distribution without memorizing the training samples [57]. Another example of a qualitative method was used by the authors of GCGAN [185], where the authors evaluated the generated data by showcasing examples


Membership Inference Attack
1
Full set of real data
GAN training set
held-out set
Accessed by attacker
2
Real training set
Undisclosed to attacker
GAN
3
Synthetic data
Accessed by attacker
4
X ∈ Training set?
Sample X from Full real set
Each sample in synthetic set
distance
5
Distance > Threshold?
Membership is inferred
(a)
Attribute Disclosure Attack
1
Real EHR Data
Protected sensitive attribute
Accessed by attacker
GAN
No direct mapping between real and synthetic records
2
Synthetic EHR data
Accessed by attacker
3
Ethnicity of K-nearest neighbor in synthetic data will help determine Ethnicity in Real
Disclosed Attributes
4
Disclosed Attributes of the real data
(b)
Model Inversion Attack
1
Real EHR Data
Protected data
2
GAN
Accessed by attacker
Synthetic EHR data
Accessed by attacker
3
01101
00101
01011
Attacker constantly queries the model, and understands how it works and its weights
4
Attacker is able to reconstruct the training data
Reconstructed Real EHR Data
(c)

Full set of real data

Patient ID	Ethnicity	Insurance Company	ICD code	....
1001	Asian	A	E05.0	......
1002	Black	B	M35.9	.....
1003	Hispanic	B	D89.3	.....
⋮	⋮	⋮	⋮	⋮

Real training set

Patient ID	Ethnicity	Insurance Company	ICD code	....
1001	Asian	A	E05.0	......
1002	Black	B	M35.9	.....

Synthetic data

Patient ID	Ethnicity	Insurance Company	ICD code	....
1001	Asian	A	E05.0	......
1832	Asian	C	M15.9	.....
1974	Hispanic	C	M19.3	.....
⋮	⋮	⋮	⋮	⋮

Real EHR Data

Patient ID	Ethnicity	Insurance company	ICD code	....
1001	Asian	A	E05.0	......
1002	Black	B	M35.9	.....
1003	Hispanic	B	D89.3	.....
⋮	⋮	⋮	⋮	⋮

Synthetic EHR data

Synthetic Patient ID	Ethnicity	Insurance Company	ICD code	....
1244	Black	B	M30.0	....
1259	Aftrican	C	I20.1	....
1276	Asian	A	E01.0	....
⋮	⋮	⋮	⋮	⋮

Disclosed Attributes

Patient ID	Ethnicity
1001	Asian
1002	Black
1003	Hispanic
⋮	

Real EHR Data

Patient ID	Ethnicity	Insurance company	ICD code	....
1001	Asian	A	E05.0	......
1002	Black	B	M35.9	.....
1003	Hispanic	B	D89.3	.....
⋮	⋮	⋮	⋮	⋮

Synthetic EHR data

Synthetic Patient ID	Ethnicity	Insurance Company	ICD code	....
1244	Black	B	M30.0	....
1259	Aftrican	C	I20.1	....
1276	Asian	A	E01.0	....
⋮	⋮	⋮	⋮	⋮

Reconstructed Real EHR Data

Patient ID	Ethnicity	Insurance company	ICD code	....
1001	Asian	A	E05.0	......
1002	Black	B	M35.9	.....
1003	Hispanic	B	D89.3	.....
⋮	⋮	⋮	⋮	⋮

Fig. 4. The major types of adversarial attacks used for empirical evaluations of GAN models. (a) Membership Inference Attack. (b) Attribute Disclosure Attack. (c) Model Inversion Attack.

where the model was able to distinguish the various disease-specific drugs from adjuvant drugs, which are based on predefined knowledge.

Researchers in machine learning often conduct ablation studies, where different components of the model are removed to evaluate the effect of the ablated component on the synthetic data. This kind of evaluation was also seen in Reference [74] to understand the role of the time-series classification layer and in Reference [113] to measure the effect of semi-supervised learning branch on the performance. In Reference [112], experiments for ablation studies are designed to evaluate the validity of network components for latent mapping and sequence generation. It is worth noting that several papers ablated various components in their model, without mentioning the "ablation studies" as seen in References [183, 183, 194].

Clinical validity and trust of the synthetic data is a major concern and a bottleneck in using synthetic data for clinical research. To address this, some papers conducted clinician evaluations, where a group of medical professionals are shown the data and asked to evaluate it based on its realism [11, 38, 111, 174]. The exact evaluation performed by clinicians can vary. For example, in References [11, 38], the clinical evaluation team was asked to give a numerical rating (from 1 to 10) of the realism of the data. Other authors asked the clinical evaluation team to classify data encounters as either real or generated and used more qualitative rating scales such as "Highly Plausible, Plausible, Implausible" [174]. The results of the clinical evaluations were then compared and reported using statistical metrics used for classification and statistical significance tasks. To measure the GAN model's ability to obey clinical constraints among variables, Constraint Violation Test was introduced where the authors calculated the differences between (max-median) and (median-min) for vital sign measurements in a tabular EHR setting [183]. The difference values on the record level were calculated, where the signs and magnitude of the difference are indicative of the constraint violations [183]. It is important to point out that the results from such qualitative techniques can be useful but are not sufficient to provide conclusive measures on the performance and quality of the GAN-based models.

6 OPEN-ACCESS DATA SOURCES

To demonstrate their usefulness for EHR-related applications, the developed GAN-based models were trained on various EHR datasets, as shown in Table 1. The datasets vary in size, openness of access, included features, and recording settings. One of the most commonly used datasets for GANs for EHRs development is **Medical Information Mart for Intensive Care (MIMIC)** III [97], which was collected from critical-care settings from Beth Israel Deaconess Medical Center in the United States [97]. Some of the included features were categorical and discrete such as demographics and patient outcomes. Others were continuous time-stamped vital-signs measurements, as well as clinical and imaging notes, and interventions. Its free access, extensive documentation, and online support community make it a suitable candidate for tabular and time-series GANs for EHRs applications. Other freely available data are *Philips eICU* [139], which is a multi-center database for critical care data. Philips eICU was collected from more than 208 hospitals throughout the United States between 2014 and 2015, making it a good choice for validating models across multi-centers. Both MIMIC and eICU datasets can be freely downloaded from *PhysioNet*, a resource that provides access to extensive collections of physiological and clinical data and related open source software [132]. Most *PhysioNet* datasets are accessible to the public users following the registration and signing of a data use agreement, with some datasets requiring additional credentialing. A recently introduced ICU dataset that was also made available on *PhysioNet* is *HiRID*, a high time-resolution ICU dataset collected from an ICU in Switzerland [89]. Similarly, other European ICU data are the Amsterdam University Medical Centers Database (AmsterdamUMCdb), which was released in 2021 [164]. HiRID and AmsterdamUMCdb are suitable datasets for critical care research for works interested in validating their models on populations outside the United States.

Another openly available data source is the **University of California Irvine (UCI)** Machine Learning Repository, which maintains 588 datasets that can be used for a wide range of applications since 2007 [8]. The repository now includes several small medical datasets. Some examples of the UCI medical datasets include the UCI breast Cancer and UCI Heart Disease datasets [8]. When using UCI datasets for benchmarking, one should be mindful of the datasets' similar names. For example, six distinct datasets include the word "breast" in their names, each with a different number of features and type of variables [8]. There is also a lack of standardization in the documentation of each of the datasets, since some have the patient identifiers and target variables included as features, while others do not. Careful and detailed documentation and reporting of the used dataset are essential to allow for accurate benchmarking and reproducibility. The development of data science competitions, such as the ones hosted on Kaggle and *PhysioNet*, resulted in open-access healthcare datasets that were used in GAN-based works such as the Kaggle Cervical Cancer and The PhysioNet Challenge 2012, as seen in Reference [99] and References [116, 117], respectively. We include the details and the links for accessing the aforementioned open-access datasets in Table 3.

A number of the surveyed works used RCT data, some of which are not directly accessible upon request and signing a user agreement. Notably, RCT datasets that have been used for several clinical research publications include **Systolic Blood Pressure Intervention Trial (SPRINT)** [71], and **Meta-Analysis Global Group in Chronic (MAGGIC)** [175], which includes data from 30 RCTs for patients with heart failure. When evaluating GANs for treatment effects, the benchmarking datasets used were the ones commonly used for causal inference applications in general. Notably, the *TWINs* [3] dataset collected for births from 1989 to 1991 in the United States was used for binary treatment research, where twins data mimic the factual and counterfactual observations for

Table 3. Summary of Open-access Datasets Commonly Used for Training Generative Models

Dataset	Patients	Encounters	Country of Origin	Multi-Center	Dataset Link
MIMIC III	38,597	49,785	United States	✗	mimic.physionet.org
Philips eICU	139,367	200,859	United States	✓	eicu-crd.mit.edu/about/eicu/
HiRID	NA	>33,000	Switzerland	✗	hirid.intensivecare.ai/
AmsterdamUMCdb	20,109	23,106	Netherlands	✗	amsterdammedicaldatascience.nl/amsterdamumcdb/
UCI Heart Disease	303	NA	United States, Hungary, & Switzerland	✓	archive.ics.uci.edu/dataset/45/heart+disease
Kaggle Cervical Cancer	3,000,000	NA	United States	✗	kaggle.com/competitions/cervical-cancer-screening/overview
PhysioNet/CinC Challenge 2012	NA	12,000	NA	NA	physionet.org/content/challenge-2012/1.0.0/

[1] NA refers to datasets with no details on the respective details evaluated in the table.

a certain outcome such as mortality within the first year of birth. Several covariates are recorded, such as race, pregnancy period, and quality of care during pregnancy. Another commonly used dataset for treatment effects is the **Infant Health and Development Program (IHDP)** data, first introduced in Reference [83], which belongs to an RCT that began in 1985 focusing on premature infants and the efficacy of educational and family support services on the infants over a 3-year period of their life [18].

Other data sources, such as **Surveillance, Epidemiology and End Results (SEER)** of the National Cancer Institute [95], Nemours Pediatrics longitudinal pediatric encounter-base data [75], and the **United Network for Organ Transplantation (UNOS)** [25], can be obtained upon request from their dedicated websites. There are several other referenced datasets in the literature; however, those were private and not accessible for open-access GANs for EHRs development.

7 FUTURE OUTLOOK

The recent developments of GANs for EHRs are promising first steps for potential research and decision support systems applications. The works we have surveyed in this article reveal many opportunities for developments in theory, algorithms, and applications. However, we believe that there are some challenges and gaps that need to be addressed and taken into consideration.

7.1 Evaluation of Synthetic Data

The lack of a universal evaluation methodology is a bottleneck in developing reliable GANs for EHRs works. As shown in Table 1, there is no standardization in the evaluation components or the metrics. Currently, researchers tend to (1) use commonly used metrics for GAN applications in other fields such as imaging and non-medical time series, (2) use metrics utilized by benchmark models, or (3) introduce their own new metrics. In addition, we noticed that the same evaluation test is referred to using different names in some cases, which adds to the confusion regarding GAN evaluation [111, 203]. When evaluating the machine learning utility, we believe that it is essential to report the results on both the synthetic and real datasets to understand the model's baseline performance and accurately determine the utility of the synthetic data for downstream tasks. We note that different metrics can lead to various limitations and tradeoffs. Therefore, currently, it is hard to determine the state-of-the-art GANs for EHRs models. While we believe that providing qualitative evaluations and analysis adds value to the studies, it is insufficient without supporting rigorous quantitative evaluations. In this work, we categorized the metrics based on the data aspect they are evaluating and whether they can be applied to each type of EHRs data. We hope that our work inspires future investigations of the newly introduced evaluation tests' strengths, limitations, and tradeoffs to standardize a guideline for selecting the metrics and their weights and matching them to the synthetic data utility.

Furthermore, in the current literature on GANs for EHRs there is no clear path to how the generated data are disseminated beyond the scope of research hypothesis testing setups. To this end, GAN training is computationally expensive and can lack stability; therefore, we recommend that future works evaluate computational complexity to allow for lightweight GAN development and dissemination.

7.2 Privacy–Similarity Tradeoff

The principles of GANs' architecture rely on the competing goals of the generator and discriminator, which overall optimize for the synthetic distribution similarity. The synthetic nature of GANs outputs implicitly preserves privacy, since there is no direct mapping between a single synthetic output and a real input. However, unintentional information leakage can ensue when dealing with sensitive information such as EHRs, as shown in the previously discussed adversarial attack mechanisms. The privacy–similarity tradeoff was a recurring theme in various works. We believe that to address the similarity–privacy tradeoff dilemma, authors should test for both factors irrespective of the chosen level of privacy guarantees. We observe that some of the early works did not consider testing for information leakage risks. Similarly, some of the works focusing on privacy improving privacy preservation of the GAN models did not adequately evaluate the data for preserving the distribution similarity. Conservative privacy guarantees such as differential privacy are helpful but can have a high cost on the fidelity and utility aspects. Considering that such strict differential privacy guarantees are not required by GDPR nor HIPPA for medical applications, we advise for at least considering one of the more relaxed privacy-preservation evaluation techniques. We believe that future directions of research should work with regulatory bodies to establish a clear guideline on the privacy risks to allow for private data owners to share synthetic data with confidence, which will open the doors for a wave of new research applications.

7.3 Generation of EHRs from Multimodal Data and Multi-Centers

The diversity resulting from the collection of various clinical information opens the doors for various research data-driven models. For example, as shown in Section 4.1, various GAN models were developed to generate different EHR data types such as tabular snapshot during patient's encounter (such as diagnosis and procedure ICD codes), as well as clinical time series collected over time (such as vital signs and laboratory measurements). However, very few works investigated generating data that captures the correlations between heterogeneous types of data, i.e., simultaneously generating EHR data with different types while modeling their underlying relationships [112]. Furthermore, even though we limited the scope of this survey to structured EHRs, in real-world applications, medical data come in other modalities, such as unstructured clinical notes and medical imaging and sensors data documented in EHR systems, which have related areas in natural language processing [145], computer vision [58], and signal processing [87] research. Leveraging information existing in EHRs with mixed modalities can help GANs generate patient records with higher fidelity. For example, the MIMIC-III Waveform database has matched physiological data that contain respiration, PPG, and ECG signals [131]. Other examples include government-based EHRs that are now integrating signals such as the NetCare EHR in Alberta, Canada. Overall, generating EHR data from a holistic perspective can also contribute to the realization of the concept of "digital twins" and personalised medicine in the future [64, 64].

Training deep neural networks requires large amounts of data that are representative of the target patient population, which usually entails training on data from multiple institutions. Despite using multiple datasets, the majority of the papers included in this work train on one dataset at a time. We believe that researchers first must overcome the challenges of feature mismatch and distribution mismatch [193] to reach the optimal application of a GAN model in different institutions. The literature is still nascent with respect to applications of GANs for EHRs for implementation in different institutions. One of the few works that explored the use of GANs for domain translation to facilitate the use of EHR data from multi-centers was *RadialGAN* [193]. Recently, the first GAN-based federated learning framework for tabular EHRs was proposed [174]. However, the authors only used a single dataset and split it into separate data silos in an experiment to simulate multi-centers. Separate GAN models were trained on each of the silos and then were later combined in a central GAN model [174]. We expect future works to investigate the feasibility and introduce new ways to implement GAN models on datasets from different institutions and explore new applications such as federated and continual learning [6] .

7.4 Reporting and Open Access Resources

Transparent reporting of training and validation datasets, preprocessing steps, hyper-parameter space, and training methodology in GAN-based applications are paramount for achieving safe use of the GAN model and for benchmarking and reproducibility of results. As noted in Reference [23], feature encoding techniques and entire hyper-parameter space are often not described despite having a substantial impact on the results for missing value imputation tasks. Without transparent and comprehensive reporting, it becomes difficult to understand the GAN model's assumptions and limitations that then impedes their safe deployment and usage.

However, we observe a positive trend of open-access work, where most of the surveyed papers published their code online. Nevertheless, some papers mentioned providing open-access code while lacking or referencing non-functional links. The open-access datasets mentioned in Section 6 allow for a wide range of GANs for EHRs applications. However, we also acknowledge that despite the usefulness of critical care and small-sized datsets in many healthcare applications, their utility is limited in some tasks. For example, generating synthetic longitudinal data is important to study prescription activities, long-term treatments effects, and other population-wide research questions. Without open-access datasets of different kinds, it will be challenging to expand GANs for EHRs research to include longitudinal data.

7.5 Integration in Clinical Applications

Using simulated data in medical practice is not new; senior academic medics compile hand-engineered simulated data to train medical students and residents as a part of their education [41]. However, using machine learning-based generated synthetic data for research and clinical support system raises concerns and questions of trust, reliability, and realism from the clinical research community. Currently, most quantitative evaluation tests and metrics are hard to interpret by medical professionals [31], which results in a gap between synthetic data and GAN development and their usage in clinical applications. To mitigate this gap, there is a need to develop evaluation tests that confirm the preservation of unique characteristics of clinical datasets that clinicians easily understand. We believe that using such metrics in conjunction with rigorous mathematical and statistical similarity evaluation will support the acceptance of the use of synthetic data. Furthermore, co-designing algorithms with clinicians generally enhances the field of machine learning to develop new architectures for various applications in healthcare.

With the increased number of works introducing new methodologies, evaluation metrics, and applications of GANs for EHRs, we believe that many of these models need to be validated on real-world large-scale EHR databases. By validating the included works on real-world EHR databases, we get a better understanding of the true scalability and reproducibility of data fidelity, utility, and privacy results. Furthermore, such validation is needed to test for the GANs' ability to capture variations of complex dependency relationships between variables stored in EHR databases from diverse clinical settings.

We believe that synthetic data have the potential to inspire a wide range of clinical research as seen in non-GAN based synthetic datasets [49, 50, 171]. With reduced time for data access and ethics approvals, as seen in Reference [73], research can be expedited, supporting the advancement of machine learning for healthcare. Overall, GANs for EHRs is a relatively new field and still has lots of capacity for improvement, especially in addressing EHR data complexity aspects such as heterogeneity, missingness, and sparsity.

REFERENCES

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. 308–318.

[2] Denis Agniel, Isaac S. Kohane, and Griffin M. Weber. 2018. Biases in electronic health record data due to processes within the healthcare system: Retrospective observational study. *Br. Med. J.* 361 (2018).

[3] Douglas Almond, Kenneth Y. Chay, and David S. Lee. 2005. The costs of low birth weight. *Quart. J. Econ.* 120, 3 (2005), 1031–1083.

[4] Martin Arjovsky and Léon Bottou. 2017. Towards principled methods for training generative adversarial networks. arXiv:1701.04862. Retrieved from https://arxiv.org/abs/1701.04862

[5] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *Proceedings of the International Conference on Machine Learning*. PMLR, 214–223.

[6] Jacob Armstrong and D Clifton. 2021. Continual learning of longitudinal health records. arXiv:2112.11944. Retrieved from https://arxiv.org/abs/2122.11944

[7] Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. 2017. Generalization and equilibrium in generative adversarial nets (gans). In *Proceedings of the International Conference on Machine Learning*. PMLR, 224–232.

[8] Arthur Asuncion and David Newman. 2007. UCI Machine Learning Repository.

[9] Gilbert Badaro, Mohammed Saeed, and Paolo Papotti. 2023. Transformers for tabular data representation: A survey of models and applications. In *Transactions of the Association for Computational Linguistics*. MIT Press, Cambridge, MA.

[10] Mrinal Kanti Baowaly, Chia-Ching Lin, Chao-Lin Liu, and Kuan-Ta Chen. 2019. Synthesizing electronic health records using improved generative adversarial networks. *J. Am. Med. Inf. Assoc.* 26, 3 (2019), 228–241.

[11] Brett K. Beaulieu-Jones, Zhiwei Steven Wu, Chris Williams, Ran Lee, Sanjeev P. Bhavnani, James Brian Byrd, and Casey S. Greene. 2019. Privacy-preserving generative deep neural networks support clinical data sharing. *Circulation: Cardiovasc. Qual. Outcomes* 12, 7 (2019), e005122.

[12] Marc G. Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. 2017. The cramer distance as a solution to biased wasserstein gradients. arXiv:1705.10743. Retrieved from https://arxiv.org/abs/1705.10743

[13] Dmitry I. Belov and Ronald D. Armstrong. 2011. Distributions of the Kullback–Leibler divergence with applications. *Br. J. Math. Statist. Psychol.* 64, 2 (2011), 291–309.

[14] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In *Noise Reduction in Speech Processing*. Springer, 1–4.

[15] Alain Berlinet and Christine Thomas-Agnan. 2011. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science & Business Media.

[16] Guthrie S. Birkhead, Michael Klompas, and Nirav R. Shah. 2015. Uses of electronic health records for public health surveillance to advance public health. *Annu. Rev. Publ. Health* 36 (2015), 345–359.

[17] Nick Black. 1996. Why we need observational studies to evaluate the effectiveness of health care. *Br. Med. J.* 312, 7040 (1996), 1215–1218.

[18] Jeanne Brooks-Gunn, Fong-ruey Liaw, and Pamela Kato Klebanov. 1992. Effects of early intervention on cognitive function of low birth weight preterm infants. *J. Pediatr.* 120, 3 (1992), 350–359.

[19] Eoin Brophy, Zhengwei Wang, Qi She, and Tomás Ward. 2023. Generative Adversarial Networks in Time Series: A Systematic Literature Review. *Comput. Surv.* 55, 10 (2023), 1–31.

[20] Dorothy Bulas and Alexia Egloff. 2013. Benefits and risks of MRI in pregnancy. In *Seminars in Perinatology*, Vol. 37. Elsevier, 301–304.

[21] Marco Caliendo and Sabine Kopeinig. 2008. Some practical guidance for the implementation of propensity score matching. *J. Econ. Surv.* 22, 1 (2008), 31–72.

[22] Ramiro Camino, Christian Hammerschmidt, and Radu State. 2018. Generating multi-categorical samples with generative adversarial networks. arXiv:1807.01202. Retrieved from https://arxiv.org/abs/1807.01202

[23] Ramiro Daniel Camino, Christian Hammerschmidt, and State Radu. 2020. Working with deep generative models and tabular data imputation.

[24] Ramiro D. Camino, Christian A. Hammerschmidt, and Radu State. 2019. Improving missing data imputation with deep generative models. arXiv:1902.10666. Retrieved from https://arxiv.org/abs/1902.10666

[25] J. Michael Cecka and Paul I. Terasaki. 1992. The UNOS scientific renal transplant registry. *Clin. Transpl.* (1992), 1–16.

[26] Taha Ceritli, Ghadeer O. Ghosheh, Vinod Kumar Chauhan, Tingting Zhu, Andrew P. Creagh, and David A. Clifton. 2023. Synthesizing mixed-type electronic health records using diffusion models. arXiv:2302.14679. Retrieved from https://arxiv.org/abs/2302.14679

[27] Tianfeng Chai and Roland R. Draxler. 2014. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.* 7, 3 (2014), 1247–1250.

[28] Krishna Chaitanya, Neerav Karani, Christian F. Baumgartner, Anton Becker, Olivio Donati, and Ender Konukoglu. 2019. Semi-supervised and task-driven data augmentation. In *International Conference on Information Processing in Medical Imaging*. Springer, 29–41.

[29] Sarath Chandar, Mitesh M. Khapra, Hugo Larochelle, and Balaraman Ravindran. 2016. Correlational neural networks. *Neural Comput.* 28, 2 (2016), 257–285.

[30] Zhengping Che, Yu Cheng, Shuangfei Zhai, Zhaonan Sun, and Yan Liu. 2017. Boosting deep learning risk prediction with generative adversarial networks for electronic health records. In *Proceedings of the IEEE International Conference on Data Mining (ICDM '17)*. IEEE, 787–792.

[31] Richard J. Chen, Ming Y. Lu, Tiffany Y. Chen, Drew F. K. Williamson, and Faisal Mahmood. 2021. Synthetic data in machine learning for medicine and healthcare. *Nat. Biomed. Eng.* (2021), 1–5.

[32] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 2180–2188.

[33] Zhaoyi Chen, Xiong Liu, William Hogan, Elizabeth Shenkman, and Jiang Bian. 2021. Applications of artificial intelligence in drug development using real-world data. *Drug Discov. Today* 26, 5 (2021), 1256–1264.

[34] Kieran Chin-Cheong, Thomas Sutter, and Julia E. Vogt. 2019. Generation of heterogeneous synthetic electronic health records using GANs. In *Workshop on Machine Learning for health (ML4H) at the 33rd Conference on Neural Information Processing Systems (NeurIPS '19)*. ETH Zurich, Institute for Machine Learning.

[35] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv:1406.1078. Retrieved from https://arxiv.org/abs/1406.1078

[36] Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. 2016. Multi-layer representation learning for medical concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1495–1504.

[37] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. 2017. Generating multi-label discrete patient records using generative adversarial networks. In *Machine Learning for Healthcare Conference*. PMLR, 286–305.

[38] Edward Choi, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. 2016. Medical concept representation learning from electronic health records and its application on heart failure prediction. arXiv:1602.03686. Retrieved from https://arxiv.org/abs/1602.03686

[39] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8789–8797.

[40] Jiebin Chu, Wei Dong, Jinliang Wang, Kunlun He, and Zhengxing Huang. 2020. Treatment effect prediction with adversarial deep learning using electronic health records. *BMC Med. Inf. Decis. Mak.* 20, 4 (2020), 1–14.

[41] Jennifer A. Cleland, Keiko Abe, and Jan-Joost Rethans. 2009. The use of simulated patients in medical education: AMEE Guide No 42. *Med. Teacher* 31, 6 (2009), 477–486.

[42] Mike Conway, Richard L. Berg, David Carrell, Joshua C. Denny, Abel N. Kho, Iftikhar J. Kullo, James G. Linneman, Jennifer A. Pacheco, Peggy Peissig, Luke Rasmussen, et al. 2011. Analyzing the heterogeneity and complexity of Electronic Health Record oriented phenotyping algorithms. In *AMIA Annual Symposium Proceedings*, Vol. 2011. American Medical Informatics Association, 274.

[43] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A. Bharath. 2018. Generative adversarial networks: An overview. *IEEE Sign. Process. Mag.* 35, 1 (2018), 53–65.

[44] Arianna Dagliati, Alberto Malovini, Valentina Tibollo, and Riccardo Bellazzi. 2021. Health informatics and EHR to support clinical research in the COVID-19 pandemic: An overview. *Brief. Bioinf.* 22, 2 (2021), 812–822.

[45] Ralph B. D'Agostino Jr. 1998. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat. Med.* 17, 19 (1998), 2265–2281.

[46] Zongyu Dai, Zhiqi Bu, and Qi Long. 2021. Multiple imputation via generative adversarial network for high-dimensional blockwise missing value problems. In *Proceedings of the 20th IEEE International Conference on Machine Learning and Applications (ICMLA '21)*. IEEE, 791–798.

[47] Zihang Dai, Zhilin Yang, Fan Yang, William W. Cohen, and Ruslan Salakhutdinov. 2017. Good semi-supervised learning that requires a bad gan. arXiv:1705.09783. Retrieved from https://arxiv.org/abs/1705.09783

[48] Saloni Dash, Andrew Yale, Isabelle Guyon, and Kristin P. Bennett. 2020. Medical time-series data generation using generative adversarial networks. In *Artificial Intelligence in Medicine: 18th International Conference on Artificial Intelligence in Medicine (AIME '20)*. Springer, 382–391.

[49] Clinical Practice Research Datalink. 2020. CPRD COVID-19 Symptoms and Risk Factors Synthetic Dataset.

[50] CPRD Cardiovascular Disease Synthetic Dataset. 2020. CPRD COVID-19 Symptoms and Risk Factors Synthetic Dataset.

[51] Chloe de Grood, Jeanna Parsons Leigh, Sean M. Bagshaw, Peter M. Dodek, Robert A. Fowler, Alan J. Forster, Jamie M. Boyd, and Henry T. Stelfox. 2018. Patient, family and provider experiences with transfers from intensive care unit to hospital ward: A multicentre qualitative study. *Can. Med. Assoc. J.* 190, 22 (2018), E669–E676.

[52] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Adv. Neural Inf. Process. Syst.* 34 (2021), 8780–8794.

[53] Josep Domingo-Ferrer, David Sánchez, and Alberto Blanco-Justicia. 2021. The limits of differential privacy (and its misuse in data release and machine learning). *Commun. ACM* 64, 7 (2021), 33–35.

[54] Christopher J. Duff, Ivonne Solis-Trapala, Owen J. Driskell, David Holland, Helen Wright, Jenna L. Waldron, Clare Ford, Jonathan J. Scargill, Martin Tran, Fahmy WF Hanna, et al. 2019. The frequency of testing for glycated haemoglobin, HbA1c, is linked to the probability of achieving target levels in patients with suboptimally controlled diabetes mellitus. *Clin. Chem. Labor. Med.* 57, 2 (2019), 296–304.

[55] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* 9, 3-4 (2014), 211–407.

[56] Khaled El Emam, Lucy Mosquera, Xi Fang, and Alaa El-Hussuna. 2022. Utility metrics for evaluating synthetic health data generation methods: Validation study. *JMIR Med. Inf.* 10, 4 (2022), e35734.

[57] Cristóbal Esteban, Stephanie L. Hyland, and Gunnar Rätsch. 2017. Real-valued (medical) time series generation with recurrent conditional gans. arXiv:1706.02633. Retrieved from https://arxiv.org/abs/1706.02633

[58] Andre Esteva, Katherine Chou, Serena Yeung, Nikhil Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric Topol, Jeff Dean, and Richard Socher. 2021. Deep learning-enabled medical computer vision. *NPJ Digit. Med.* 4, 1 (2021), 1–9.

[59] Renee Fekieta, Alana Rosenberg, Beth Hodshon, Shelli Feder, Sarwat I. Chaudhry, and Beth L. Emerson. 2021. Organisational factors underpinning intra-hospital transfers: A guide for evaluating context in quality improvement. *Health Syst.* 10, 4 (2021), 239–248.

[60] Tanner Fiez, Benjamin Chasnov, and Lillian J. Ratliff. 2019. Convergence of learning dynamics in stackelberg games. arXiv:1906.01217. Retrieved from https://arxiv.org/abs/1906.01217

[61] Centers for Disease Control, Prevention, et al. 2003. HIPAA privacy rule and public health. Guidance from CDC and the US Department of Health and Human Services. *Morbid. Mortal. Week. Rep.* 52, Suppl 1 (2003), 1–17.

[62] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. 1322–1333.

[63] Qiyang Ge, Xuelin Huang, Shenying Fang, Shicheng Guo, Yuanyuan Liu, Wei Lin, and Momiao Xiong. 2020. Conditional generative adversarial networks for individualized treatment effect estimation and treatment selection. *Front. Genet.* 11 (2020).

[64] Jeremy Georges-Filteau and Elisa Cirillo. 2020. Synthetic observational health data with GANs: From slow adoption to a boom in medical research and ultimately digital twins? arXiv:2005.13510. Retrieved from https://arxiv.org/abs/2005.13510

[65] Shantanu Ghosh, Christina Boucher, Jiang Bian, and Mattia Prosperi. 2021. Propensity score synthetic augmentation matching using generative adversarial networks (PSSAM-GAN). *Comput. Methods Progr. Biomed. Update* 1 (2021), 100020.

[66] Cory E. Goldstein, Charles Weijer, Jamie C. Brehaut, Dean A. Fergusson, Jeremy M. Grimshaw, Austin R. Horn, and Monica Taljaard. 2018. Ethical issues in pragmatic randomized controlled trials: A review of the recent literature identifies gaps in ethical argumentation. *BMC Med. Ethics* 19, 1 (2018), 1–10.

[67] Andre Goncalves, Priyadip Ray, Braden Soper, Jennifer Stevens, Linda Coyle, and Ana Paula Sales. 2020. Generation and evaluation of synthetic patient data. *BMC Med. Res. Methodol.* 20, 1 (2020), 1–40.

[68] Ian Goodfellow. 2016. Nips 2016 tutorial: Generative adversarial networks. arXiv:1701.00160. Retrieved from https://arxiv.org/abs/1701.00160

[69] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial networks. *Adv. Neural Inf. Process. Syst.* 27 (2014), 2672–2680.

[70] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *J. Mach. Learn. Res.* 13, 1 (2012), 723–773.

[71] SPRINT Research Group. 2015. A randomized trial of intensive versus standard blood-pressure control. *New Engl. J. Med.* 373, 22 (2015), 2103–2116.

[72] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. 2017. Improved training of wasserstein gans. arXiv:1704.00028. Retrieved from http://arxiv.org/abs/1704.00058

[73] Aixia Guo, Randi E. Foraker, Robert M. MacGregor, Faraz M. Masood, Brian P. Cupps, and Michael K. Pasque. 2020. The use of synthetic electronic health record data and deep learning to improve timing of high-risk heart failure surgical intervention by predicting proximity to catastrophic decompensation. *Front. Digit. Health* (2020), 44.

[74] Mehak Gupta, Thao-Ly T Phan, H Timothy Bunnell, and Rahmatollah Beheshti. 2021. Concurrent imputation and prediction on EHR data using Bi-Directional GANs: Bi-GANs for EHR imputation and prediction. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics.* 1–9.

[75] Mehak Gupta, Thao-Ly T Phan, Timothy Bunnell, and Rahmatollah Beheshti. 2019. Obesity prediction with EHR data: A deep learning approach with interpretable elements. arXiv:1912.02655. Retrieved from https://arxiv.org/abs/1912.02655

[76] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. 2019. Logan: Membership inference attacks against generative models. In *Proceedings on Privacy Enhancing Technologies (PoPETs '19)*, Vol. 2019. De Gruyter, 133–152.

[77] RJ Hayes and S Bennett. 1999. Simple sample size calculation for cluster-randomized trials. *Int. J. Epidemiol.* 28, 2 (1999), 319–326.

[78] Catherine Helmer, Karine Pérès, Antoine Pariente, Florence Pasquier, Sophie Auriacombe, Michel Poncet, Florence Portet, Olivier Rouaud, Karen Ritchie, Christophe Tzourio, et al. 2008. Primary and secondary care consultations in elderly demented individuals in France. *Dementia Geriatr. Cogn. Disord.* 26, 5 (2008), 407–415.

[79] J. Henry, Yuriy Pylypchuk, Talisha Searcy, and Vaishali Patel. 2016. Adoption of electronic health record systems among US non-federal acute care hospitals: 2008–2015. *ONC Data Brief* 35 (2016), 1–9.

[80] Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. 2022. Synthetic data generation for tabular health records: A systematic review. *Neurocomputing* (2022).

[81] Emily Herrett, Arlene M. Gallagher, Krishnan Bhaskaran, Harriet Forbes, Rohini Mathur, Tjeerd Van Staa, and Liam Smeeth. 2015. Data resource profile: Clinical practice research datalink (CPRD). *Int. J. Epidemiol.* 44, 3 (2015), 827–836.

[82] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2016. beta-vae: Learning basic visual concepts with a constrained variational framework.

[83] Jennifer L. Hill. 2011. Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Stat.* 20, 1 (2011), 217–240.

[84] R. Devon Hjelm, Athul Paul Jacob, Tong Che, Adam Trischler, Kyunghyun Cho, and Yoshua Bengio. 2017. Boundary-seeking generative adversarial networks. arXiv:1702.08431. Retrieved from https://arxiv.org/abs/1702.08431

[85] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* 33 (2020), 6840–6851.

[86] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9, 8 (1997), 1735–1780.

[87] Mohammad-Parsa Hosseini, Amin Hosseini, and Kiarash Ahi. 2020. A review on machine learning for EEG signal processing in bioengineering. *IEEE Rev. Biomed. Eng.* 14 (2020), 204–218.

[88] Mohammad Hossin and Md Nasir Sulaiman. 2015. A review on evaluation metrics for data classification evaluations. *Int. J. Data Min. Knowl. Manage. Process* 5, 2 (2015), 1.

[89] Stephanie L. Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, et al. 2020. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nat. Med.* 26, 3 (2020), 364–373.

[90] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2017. Globally and locally consistent image completion. *ACM Trans. Graph.* 36, 4 (2017), 1–14.

[91] Michael Imhoff. 1992. Acquisition of ICU data: Concepts and demands. *Int. J. Clin. Monitor. Comput.* 9, 4 (1992), 229–237.

[92] R. Indhumathi and S. Sathiya Devi. 2021. Healthcare Cramér generative adversarial network (HCGAN). *Distrib. Parallel Datab.* (2021), 1–17.

[93] Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. arXiv:1611.01144.

[94] Nathalie Japkowicz and Shaju Stephen. 2002. The class imbalance problem: A systematic study. *Intell. Data Anal.* 6, 5 (2002), 429–449.

[95] Ahmedin Jemal, Andrea Thomas, Taylor Murray, Michael Thun, et al. 2002. Cancer statistics, 2002. *Ca Cancer J. Clin.* 52, 1 (2002), 23–47.

[96] Jiwoong J. Jeong, Amara Tariq, Tobiloba Adejumo, Hari Trivedi, Judy W. Gichoya, and Imon Banerjee. 2022. Systematic review of generative adversarial networks (GANs) for medical image classification and segmentation. *J. Digit. Imag.* (2022), 1–16.

[97] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Sci. Data* 3, 1 (2016), 1–9.

[98] James Jordon, Alan Wilson, and Mihaela van der Schaar. 2020. Synthetic data: Opening the data floodgates to enable faster, more directed development of machine learning methods. arXiv:2012.04580. Retrieved from https://arxiv.org/abs/2012.04580

[99] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. 2018. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*.

[100] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive growing of gans for improved quality, stability, and variation. arXiv:1710.10196. Retrieved from https://arxiv.org/abs/1710.10196

[101] Ismail Keshta and Ammar Odeh. 2021. Security and privacy of electronic health records: Concerns and challenges. *Egypt. Inf. J.* 22, 2 (2021), 177–183.

[102] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. 2022. Transformers in vision: A survey. *ACM Comput. Surv.* 54, 10s (2022), 1–41.

[103] Ellen Kim, Samuel M. Rubinstein, Kevin T. Nead, Andrzej P. Wojcieszynski, Peter E. Gabriel, and Jeremy L. Warner. 2019. The evolving use of electronic health records (EHR) for research. In *Seminars in Radiation Oncology*, Vol. 29. Elsevier, 354–361.

[104] Joanne Kim, Gilad Horowitz, Michael Hong, Mario Orsini, Sylvia L. Asa, and Kevin Higgins. 2017. The dangers of parathyroid biopsy. *J. Otolaryngol. Head Neck Surg.* 46, 1 (2017), 1–4.

[105] Diederik P. Kingma and Max Welling. 2013. Auto-encoding variational bayes. arXiv:1312.6114. Retrieved from https://arxiv.org/abs/1312.6114

[106] Virginia Klema and Alan Laub. 1980. The singular value decomposition: Its computation and some applications. *IEEE Trans. Automat. Contr.* 25, 2 (1980), 164–176.

[107] Andrew W. Knight and Timothy P. Senior. 2006. The common problem of rare disease in general practice. *Med. J. Austr.* 185, 2 (2006), 82–83.

[108] David M. Kreindler and Charles J. Lumsden. 2006. The effects of the irregular sample and missing data in time series analysis. *Nonlinear Dynam. Psychol. Life Sci.* (2006).

[109] John Krumm. 2018. *Ubiquitous Computing Fundamentals*. CRC Press.

[110] The Lancet. 2018. Personalised medicine in the UK. *Lancet (London, Engl.)* 391, 10115 (2018), e1.

[111] Dongha Lee, Hwanjo Yu, Xiaoqian Jiang, Deevakar Rogith, Meghana Gudala, Mubeen Tejani, Qiuchen Zhang, and Li Xiong. 2020. Generating sequential electronic health records using dual adversarial autoencoder. *J. Am. Med. Inf. Assoc.* 27, 9 (2020), 1411–1419.

[112] Jin Li, Benjamin J. Cairns, Jingsong Li, and Tingting Zhu. 2021. Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. arXiv:2112.12047. Retrieved from https://arxiv.org/abs/2112.12047

[113] Wenyuan Li, Yunlong Wang, Yong Cai, Corey Arnold, Emily Zhao, and Yilian Yuan. 2018. Semi-supervised rare disease detection using generative adversarial network. arXiv:1812.00547. Retrieved from https://arxiv.org/abs/1812.00547

[114] Yunzhe Li, Kun Kuang, Bo Li, Peng Cui, Jianrong Tao, Hongxia Yang, and Fei Wu. 2020. Continuous treatment effect estimation via generative adversarial de-confounding. In *Proceedings of the KDD Workshop on Causal Discovery*. PMLR, 4–22.

[115] Roderick J. A. Little and Donald B. Rubin. 2019. *Statistical Analysis with Missing Data*. Vol. 793. John Wiley & Sons.

[116] Yonghong Luo, Xiangrui Cai, Ying Zhang, Jun Xu, et al. 2018. Multivariate time series imputation with generative adversarial networks. *Adv. Neural Inf. Process. Syst.* 31 (2018).

[117] Yonghong Luo, Ying Zhang, Xiangrui Cai, and Xiaojie Yuan. 2019. E2gan: End-to-end generative adversarial network for multivariate time series imputation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 3094–3100.

[118] Jeanne M. Madden, Matthew D. Lakoma, Donna Rusinak, Christine Y. Lu, and Stephen B. Soumerai. 2016. Missing clinical and behavioral health data in a large electronic health record (EHR) system. *J. Am. Med. Inf. Assoc.* 23, 6 (2016), 1143–1149.

[119] A. P. Majtey, P. W. Lamberti, and D. P. Prato. 2005. Jensen-Shannon divergence as a measure of distinguishability between mixed quantum states. *Phys. Rev. A* 72, 5 (2005), 052310.

[120] Tatiana Malygina, Elena Ericheva, and Ivan Drokin. 2019. Data augmentation with GAN: Improving chest X-Ray pathologies prediction on class-imbalanced cases. In *International Conference on Analysis of Images, Social Networks and Texts*. Springer, 321–334.

[121] Navin Kumar Manaswi. 2018. Rnn and lstm. In *Deep Learning with Applications Using Python*. Springer, 115–126.

[122] Benjamin Marlin. 2008. *Missing Data Problems in Machine Learning*. Ph. D. Dissertation.

[123] Frank J. Massey Jr. 1951. The Kolmogorov-Smirnov test for goodness of fit. *J. Am. Statist. Assoc.* 46, 253 (1951), 68–78.

[124] Stan Matwin, Jordi Nin, Morvarid Sehatkar, and Tomasz Szapiro. 2015. A review of attribute disclosure control. *Adv. Res. Data Priv.* (2015), 41–61.

[125] Matthew McDermott, Tom Yan, Tristan Naumann, Nathan Hunt, Harini Suresh, Peter Szolovits, and Marzyeh Ghassemi. 2018. Semi-supervised biomedical translation with cycle wasserstein regression GANs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[126] Roseanne McNamee. 2003. Confounding and confounders. *Occupat. Environ. Med.* 60, 3 (2003), 227–234.

[127] Bradley D. Menz, Sophie L. Stocker, Nick Verougstraete, Danijela Kocic, Peter Galettis, Christophe P. Stove, and Stephanie E. Reuter. 2021. Barriers and opportunities for the clinical implementation of therapeutic drug monitoring in oncology. *Br. J. Clin. Pharmacol.* 87, 2 (2021), 227–236.

[128] Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. 2016. Unrolled generative adversarial networks. arXiv:1611.02163. Retrieved from https://arxiv.org/abs/1611.02163

[129] Lu Mi, Macheng Shen, and Jingzhao Zhang. 2018. A probe towards understanding gan and vae models. arXiv:1812.05676. Retrieved from https://arxiv.org/abs/1812.05676

[130] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. arXiv:1411.1784. Retrieved from https://arxiv.org/abs/1411.1784

[131] Benjamin Moody, George Moody, Mauricio Villarroel, Gari D. Clifford, and Ikaro Silva. 2020. MIMIC-III Waveform Database Matched Subset. Retrieved from https://physionet.org/content/mimic3wdb-matched/1.0/

[132] George B. Moody, Roger G. Mark, and Ary L. Goldberger. 2001. PhysioNet: A web-based resource for the study of physiologic signals. *IEEE Eng. Med. Biol. Mag.* 20, 3 (2001), 70–75.

[133] Sumit Mukherjee, Yixi Xu, Anusua Trivedi, and Juan Lavista Ferres. 2020. Protecting GANs against privacy attacks by preventing overfitting.

[134] Fulufhelo V. Nelwamondo, Shakir Mohamed, and Tshilidzi Marwala. 2007. Missing data: A comparison of neural network and expectation maximization techniques. *Curr. Sci.* (2007), 1514–1521.

[135] Simon J. Newsome, Ruth H. Keogh, and Rhian M. Daniel. 2018. Estimating long-term treatment effects in observational data: A comparison of the performance of different methods under real-world uncertainty. *Stat. Med.* 37, 15 (2018), 2367–2390.

[136] Frank Nielsen. 2020. On a generalization of the Jensen–Shannon divergence and the Jensen–Shannon centroid. *Entropy* 22, 2 (2020), 221.

[137] Augustus Odena. 2016. Semi-supervised learning with generative adversarial networks. arXiv:1606.01583. Retrieved from https://arxiv.org/abs/1606.01583

[138] Nicolas Papernot, Martn Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. 2016. Semi-supervised knowledge transfer for deep learning from private training data. arXiv:1610.05755. Retrieved from https://arxiv.org/abs/1610.05755

[139] Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. 2018. The eICU collaborative research database, a freely available multi-center database for critical care research. *Sci. Data* 5, 1 (2018), 1–13.

[140] Yeping Lina Qiu, Hong Zheng, and Olivier Gevaert. 2020. Genomic data imputation with variational auto-encoders. *GigaScience* 9, 8 (2020).

[141] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv:1511.06434. Retrieved from https://arxiv.org/abs/1511.06434

[142] Sina Rashidian, Fusheng Wang, Richard Moffitt, Victor Garcia, Anurag Dutt, Wei Chang, Vishwam Pandya, Janos Hajagos, Mary Saltz, and Joel Saltz. 2020. SMOOTH-GAN: Towards sharp and smooth synthetic EHR data generation. In *Artificial Intelligence in Medicine: 18th International Conference on Artificial Intelligence in Medicine (AIME '20)*. Springer, 37–48.

[143] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. 2019. Generating diverse high-fidelity images with vq-vae-2. *Advances in Neural Information Processing Systems* 32 (2019).

[144] Paul R. Rosenbaum. 2002. Covariance adjustment in randomized experiments and observational studies. *Statist. Sci.* 17, 3 (2002), 286–327.

[145] Halley Ruppel, Aashish Bhardwaj, Raj N. Manickam, Julia Adler-Milstein, Marc Flagg, Manuel Ballesca, and Vincent X. Liu. 2020. Assessment of electronic health record search patterns and practices by practitioners in a large integrated health care system. *JAMA Netw. Open* 3, 3 (2020), e200512–e200512.

[146] Ludger Rüschendorf. 1985. The Wasserstein distance and approximation theorems. *Probab. Theory Relat. Fields* 70, 1 (1985), 117–129.

[147] Pegah Salehi, Abdolah Chalechale, and Maryam Taghizadeh. 2020. Generative adversarial networks (GANs): An overview of theoretical model, evaluation metrics, and recent developments. arXiv:2005.13178. Retrieved from https://arxiv.org/abs/2005.13178

[148] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Adv. Neural Inf. Process. Syst.* 29 (2016), 2234–2242.

[149] Sergio Sanchez-Martinez, Oscar Camara, Gemma Piella, Maja Cikes, Miguel Angel Gonzalez Ballester, Marius Miron, Alfredo Vellido, Emilia Gomez, Alan Fraser, and Bart Bijnens. 2019. Machine learning for clinical decision-making: Challenges and opportunities.

[150] Robert William Sanson-Fisher, Billie Bonevski, Lawrence W. Green, and Cate D'Este. 2007. Limitations of the randomized controlled trial in evaluating population-based health interventions. *Am. J. Prevent. Med.* 33, 2 (2007), 155–161.

[151] Sara Santiso, Arantza Casillas, and Alicia Pérez. 2019. The class imbalance problem detecting adverse drug reactions in electronic health records. *Health Inf. J.* 25, 4 (2019), 1768–1778.

[152] Ashish Sarraju, Andrew Ward, Sukyung Chung, Jiang Li, David Scheinker, and Fàtima Rodriguez. 2021. Machine learning approaches improve risk stratification for secondary cardiovascular disease prevention in multiethnic patients. *Open Heart* 8, 2 (2021), e001802.

[153] Uri Shalit, Fredrik D. Johansson, and David Sontag. 2017. Estimating individual treatment effect: Generalization bounds and algorithms. In *International Conference on Machine Learning*. PMLR, 3076–3085.

[154] Farah Shamout, Tingting Zhu, and David A. Clifton. 2020. Machine learning for clinical outcome prediction. *IEEE Rev. Biomed. Eng.* 14 (2020), 116–126.

[155] A. Mi Shin, In Hee Lee, Gyeong Ho Lee, Hee Joon Park, Hyung Seop Park, Kyung Il Yoon, Jung Jeung Lee, and Yoon Nyun Kim. 2010. Diagnostic analysis of patients with essential hypertension using association rule mining. *Healthcare Inf. Res.* 16, 2 (2010), 77–81.

[156] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *Proceedings of the IEEE Symposium on Security and Privacy (SP '17)*. IEEE, 3–18.

[157] Leif Sörnmo and Pablo Laguna. 2006. Electrocardiogram (ECG) signal processing. In *Wiley Encyclopedia of Biomedical Engineering*.

[158] Daniel J. Stekhoven and Peter Bühlmann. 2012. MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28, 1 (2012), 112–118.

[159] Chenxi Sun, Shenda Hong, Moxian Song, and Hongyan Li. 2020. A review of deep learning methods for irregularly sampled medical time series data. arXiv:2010.12493. Retrieved from https://arxiv.org/abs/2010.12493

[160] Reed T. Sutton, David Pincock, Daniel C. Baumgart, Daniel C. Sadowski, Richard N. Fedorak, and Karen I. Kroeker. 2020. An overview of clinical decision support systems: Benefits, risks, and strategies for success. *NPJ Digit. Med.* 3, 1 (2020), 1–10.

[161] Nicholas P. Tatonetti, J. C. Denny, S. N. Murphy, G. H. Fernald, G. Krishnan, V. Castro, P. Yue, P. S. Tsau, I. Kohane, D. M. Roden, et al. 2011. Detecting drug interactions from adverse-event reports: Interaction between paroxetine and pravastatin increases blood glucose levels. *Clin. Pharmacol. Therapeut.* 90, 1 (2011), 133–142.

[162] Jonathan M. Teich. 1998. Information systems support for emergency medicine. *Ann. Emerg. Med.* 31, 3 (1998), 304–307.

[163] Lucas Theis, Aäron van den Oord, and Matthias Bethge. 2015. A note on the evaluation of generative models. arXiv:1511.01844. Retrieved from https://arxiv.org/abs/1511.01844

[164] Patrick J. Thoral, Jan M. Peppink, Ronald H. Driessen, Eric J. G. Sijbrands, Erwin J. O. Kompanje, Lewis Kaplan, Heatherlee Bailey, Jozef Kesecioglu, Maurizio Cecconi, Matthew Churpek, et al. 2021. Sharing ICU patient data responsibly under the society of critical care medicine/european society of intensive care medicine joint data science collaboration: The amsterdam university medical centers database (AmsterdamUMCdb) example. *Crit. Care Med.* 49, 6 (2021), e563.

[165] Amirsina Torfi and Edward A. Fox. 2020. Cor-gan: Correlation-capturing convolutional neural networks for generating synthetic healthcare records.

[166] Dmitrii Usynin, Alexander Ziller, Marcus Makowski, Rickmer Braren, Daniel Rueckert, Ben Glocker, Georgios Kaissis, and Jonathan Passerat-Palmbach. 2021. Adversarial interference and its mitigations in privacy-preserving collaborative machine learning. *Nat. Mach. Intell.* 3, 9 (2021), 749–758.

[167] SS Vallender. 1974. Calculation of the Wasserstein distance between probability distributions on the line. *Theory Probab. Appl.* 18, 4 (1974), 784–786.

[168] Stef Van Buuren and Karin Groothuis-Oudshoorn. 2011. mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* 45 (2011), 1–67.

[169] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[170] Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). In *A Practical Guide*, 1st ed. Springer International, Cham.

[171] Jason Walonoski, Sybil Klaus, Eldesia Granger, Dylan Hall, Andrew Gregorowicz, George Neyarapally, Abigail Watson, and Jeff Eastman. 2020. Synthea™ Novel coronavirus (COVID-19) model and synthetic data set. *Intell.-Bas. Med.* 1 (2020), 100007.

[172] Lu Wang, Wei Zhang, and Xiaofeng He. 2019. Continuous patient-centric sequence generation via sequentially coupled adversarial learning. In *Database Systems for Advanced Applications*, Guoliang Li, Jun Yang, Joao Gama, Juggapong Natwichai, and Yongxin Tong (Eds.). Springer International Publishing, Cham, 36–52.

[173] Shuo Wang, Carsten Rudolph, Surya Nepal, Marthie Grobler, and Shangyu Chen. 2020. PART-GAN: Privacy-preserving time-series sharing. In *International Conference on Artificial Neural Networks*. Springer, 578–593.

[174] John Weldon, Tomas Ward, and Eoin Brophy. 2021. Generation of synthetic electronic health records using a federated GAN. arXiv:2109.02543. Retrieved from https://arxiv.org/abs/2109.02543

[175] Chih M. Wong, Nathaniel M. Hawkins, Mark C. Petrie, Pardeep S. Jhund, Roy S. Gardner, Cono A. Ariti, Katrina K. Poppe, Nikki Earle, Gillian A. Whalley, Iain B. Squire, et al. 2014. Heart failure in younger patients: The Meta-analysis Global Group in Chronic Heart Failure (MAGGIC). *Eur. Heart J.* 35, 39 (2014), 2714–2721.

[176] Mi-Ja Woo, Jerome P. Reiter, Anna Oganian, and Alan F. Karr. 2009. Global measures of data utility for microdata masked for disclosure limitation. *J. Priv. Confident.* 1, 1 (2009).

[177] Cao Xiao, Edward Choi, and Jimeng Sun. 2018. Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review. *J. Am. Med. Inf. Assoc.* 25, 10 (2018), 1419–1428.

[178] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. 2018. Differentially private generative adversarial network. arXiv:1802.06739. Retrieved from https://arxiv.org/abs/1802.06739

[179] Pranjul Yadav, Michael Steinbach, Vipin Kumar, and Gyorgy Simon. 2018. Mining electronic health records (EHRs) A survey. *ACM Comput. Surv.* 50, 6 (2018), 1–40.

[180] Alexandre Yahi, Rami Vanguri, Noémie Elhadad, and Nicholas P. Tatonetti. 2017. Generative adversarial networks for electronic health records: A framework for exploring and evaluating methods for predicting drug-induced laboratory test trajectories. arXiv:1712.00164. Retrieved from https://arxiv.org/abs/1712.00164

[181] Andrew Yale, Saloni Dash, Karan Bhanot, Isabelle Guyon, John S. Erickson, and Kristin P. Bennett. 2020. Synthesizing quality open data assets from private health research studies. In *Business Information Systems Workshops: BIS '20 International Workshops, Revised Selected Papers 23*. Springer, 324–335.

[182] Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P. Bennett. 2020. Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing* 416 (2020), 244–255.

[183] Chao Yan, Ziqi Zhang, Steve Nyemba, and Bradley A. Malin. 2020. Generating electronic health records with multiple data types and constraints. In *AMIA Annual Symposium Proceedings*, Vol. 2020. American Medical Informatics Association, 1335.

[184] Dingdong Yang, Seunghoon Hong, Yunseok Jang, Tianchen Zhao, and Honglak Lee. 2019. Diversity-sensitive conditional generative adversarial networks. arXiv:1901.09024. Retrieved from https://arxiv.org/abs/1901.09024

[185] Fan Yang, Zhongping Yu, Yunfan Liang, Xiaolu Gan, Kaibiao Lin, Quan Zou, and Yifeng Zeng. 2019. Grouped correlational generative adversarial networks for discrete electronic health records. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM '19)*. IEEE, 906–913.

[186] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2022. Diffusion models: A comprehensive survey of methods and applications. arXiv:2209.00796. Retrieved from https://arxiv.org/abs/2209.00796

[187] Yun Yang, Fengtao Nan, Po Yang, Qiang Meng, Yingfu Xie, Dehai Zhang, and Khan Muhammad. 2019. GAN-based semi-supervised learning approach for clinical decision support in health-IoT platform. *IEEE Access* 7 (2019), 8048–8057.

[188] Yinchong Yang, Zhiliang Wu, Volker Tresp, and Peter A. Fasching. 2019. Categorical EHR imputation with generative adversarial nets. In *Proceedings of the IEEE International Conference on Healthcare Informatics (ICHI '19)*. IEEE, 1–10.

[189] Xin Yi and Paul Babyn. 2018. Sharpness-aware low-dose CT denoising using conditional generative adversarial network. *J. Digit. Imag.* 31, 5 (2018), 655–669.

[190] Jinsung Yoon, Lydia N. Drumright, and Mihaela Van Der Schaar. 2020. Anonymization through data synthesis using generative adversarial networks (ads-gan). *IEEE J. Biomed. Health Inf.* 24, 8 (2020), 2378–2388.

[191] Jinsung Yoon, Daniel Jarrett, and Mihaela Van der Schaar. 2019. Time-series generative adversarial networks. *Adv. Neural Inf. Process. Syst.* 32 (2019).

[192] Jinsung Yoon, James Jordon, and Mihaela Schaar. 2018. GAIN: Missing data imputation using generative adversarial nets. In *International Conference on Machine Learning*. PMLR, 5689–5698.

[193] Jinsung Yoon, James Jordon, and Mihaela Schaar. 2018. RadialGAN: Leveraging multiple datasets to improve target-specific predictive models using Generative Adversarial Networks. In *International Conference on Machine Learning*. PMLR, 5699–5707.

[194] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. 2018. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*.

[195] Michael P. Young, Valerie J. Gooder, Karen McBride, Brent James, and Elliott S. Fisher. 2003. Inpatient transfers to the intensive care unit. *J. Gen. Internal Med.* 18, 2 (2003), 77–83.

[196] Kezi Yu, Yunlong Wang, Yong Cai, Cao Xiao, Emily Zhao, Lucas Glass, and Jimeng Sun. 2019. Rare disease detection by sequence modeling with generative adversarial networks. arXiv:1907.01022. Retrieved from https://arxiv.org/abs/1907.01022

[197] Shuangfei Zhai, Yu Cheng, Rogerio Feris, and Zhongfei Zhang. 2016. Generative adversarial networks as variational training of energy based models. arXiv:1611.01799. Retrieved from https://arxiv.org/abs/1611.01799

[198] Chengqi Zhang and Shichao Zhang. 2003. *Association Rule Mining: Models and Algorithms.* Vol. 2307. Springer.

[199] Da Zhang, Ming Ma, and Likun Xia. 2022. A comprehensive review on GANs for time-series signals. *Neural Comput. Appl.* (2022), 1–21.

[200] Hongyang Zhang and David P. Woodruff. 2018. Medical missing data imputation by stackelberg GAN. Carnegie Mellon University.

[201] Yue Zhang, Shun Miao, Tommaso Mansi, and Rui Liao. 2020. Unsupervised X-ray image segmentation with task driven generative adversarial networks. *Med. Image Anal.* 62 (2020), 101664.

[202] Zhifei Zhang, Yang Song, and Hairong Qi. 2018. Decoupled learning for conditional adversarial networks. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV '18).* IEEE, 700–708.

[203] Ziqi Zhang, Chao Yan, Thomas A. Lasko, Jimeng Sun, and Bradley A. Malin. 2021. SynTEG: A framework for temporal structured electronic health data simulation. *J. Am. Med. Inf. Assoc.* 28, 3 (2021), 596–604.

[204] Ziqi Zhang, Chao Yan, Diego A. Mesa, Jimeng Sun, and Bradley A. Malin. 2020. Ensuring electronic medical record simulation through better training, modeling, and evaluation. *J. Am. Med. Inf. Assoc.* 27, 1 (2020), 99–108.

[205] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision.* 2223–2232.

[206] Xiaojin Zhu and Andrew B. Goldberg. 2009. Introduction to semi-supervised learning. *Synth. Lect. Artif. Intell. Mach. Learn.* 3, 1 (2009), 1–130.

Received 14 December 2022; revised 24 September 2023; accepted 1 December 2023